

September 1993

LIDS-TH-2195

Research Supported By:

National Science Foundation

NDSEG Fellowships

A Linguistic Feature Representation of the Speech Waveform

Ellen Marie Eide

September 1993

LIDS-TH-2195

Sponsor Acknowledgments

National Science Foundation

NDSEG Fellowships

A Linguistic Feature Representation of the Speech Waveform

Ellen Marie Eide

This report is based on the unaltered thesis of Ellen Marie Eide submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the Massachusetts Institute of Technology in September 1993.

This research was conducted at the M.I.T. Laboratory for Information and Decision Systems with research support gratefully acknowledged by the above mentioned sponsors.

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

**A Linguistic Feature Representation of the
Speech Waveform**

by

Ellen Marie Eide

Submitted to the Department of Electrical Engineering and
Computer Science

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1993

© Massachusetts Institute of Technology 1993. All rights reserved.

Author 
Department of Electrical Engineering and Computer Science
August 19, 1993

Certified by 
Sanjoy K. Mitter
Professor of Electrical and Computer Engineering
Thesis Supervisor

Accepted by
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

A Linguistic Feature Representation of the Speech Waveform

by

Ellen Marie Eide

Submitted to the Department of Electrical Engineering and Computer Science
on August 20, 1993, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Linguistic theory views a phoneme as a shorthand notation for a set of features which describe the operations of the articulators required to produce the sound.

In this thesis, a representation of the speech waveform in terms of the same set of distinctive linguistic features as is used to describe the abstract phonemes is developed. The ultimate goal of such a representation is robust lexical access in the task of continuous speech recognition.

The algorithm which results in the feature representation of the waveform proceeds hierarchically. The first stage of the processing estimates the broad class of speech sounds most-likely represented by each frame in the waveform. Subsequent processing, which is dependent on the estimated broad class, proceeds in terms of local and more global queries as to the feature composition in the neighborhood of each frame. Various modeling choices within the general framework are possible.

The fidelity of the representation is tested on the tasks of phoneme classification and recognition.

Thesis Supervisor: Sanjoy K. Mitter

Title: Professor of Electrical and Computer Engineering

Acknowledgments

The synthesis of this document represents the combined efforts of a number of people, especially the members of my thesis committee. I thank my advisor, Sanjoy Mitter, whose remarkable ability to ask the correct questions has shaped my research and honed my analytic skills. The other members of the committee, Ken Stevens and Lou Braida, have also contributed immeasurably to the work through patient discussions of the procedures and their relationship to speech production and perception.

In addition to the thesis committee, the work reflects the input of a number of researchers at BBN. Thanks especially to Robin Rohlicek and Herb Gish for all their help along the way.

Finally, thanks to my family and friends for their support and encouragement over the past five years.

The academic work was sponsored by NSF and NDSEG fellowships.

Contents

1. Introduction	10
1.1 The Goal and Its Motivation	10
1.2 Terminology	13
1.3 Structure of the Thesis	14
1.4 Experimental Data	15
1.4.1 Data Partitioning	16
1.4.2 Waveform Representation	16
1.4.3 Interpretation of Labels	17
2. A Distinctive Feature Representation of Phonemes	19
2.1 Introduction	19
2.2 A Linguistic Feature Set	20
2.3 Feature Configurations of Phonemes	21
3. A Distinctive Feature Representation – Local Processing	27
3.1 Overview of the Procedure	27
3.2 Gaussian Models for Broad Class Estimation	28
3.3 Estimating Broad Classes	31
3.4 Broad Class Accuracy	33
3.5 Gaussian Models for Linguistic Features	34
3.6 Dependent Modeling of Place Features	36
3.7 Estimating Linguistic Features	37

3.8	Results of Linguistic Feature Estimation	38
3.8.1	Method of Analysis	39
3.8.2	<i>d'</i> Analysis of Local Processing	41
3.9	Weaknesses of the Local Processing Algorithm	45
4.	A Distinctive Feature Representation – Global Processing	46
4.1	Introduction	46
4.2	Transition Modeling	47
4.3	Dimension Reduction	48
4.4	Probability of Features from Transition Models	49
4.5	Results of Global Processing	50
5.	Phoneme Classification and Recognition	61
5.1	Phoneme Classification	61
5.1.1	Introduction	61
5.1.2	Baseline Experiment: Gaussian Models	62
5.1.3	Phoneme Classification from Feature Probabilities	63
5.1.4	Results of Phoneme Classification	64
5.2	Phoneme Recognition	68
5.2.1	Procedure	68
5.2.2	Scoring	69
5.2.3	Results of Phoneme Recognition	70
6.	Discussion	71
6.1	Speech in the Context of General Pattern Recognition	71
6.2	The Use of Non-Linguistically-Motivated Features	72
6.3	Potential Improvements Within The Existing Framework	72
6.3.1	Robust Global Processing	74
6.3.2	Combining Local and Global Information	75
6.3.3	Database Labels	76

6.4	A Potential Improvement Through Feedback	77
6.5	Summary	78
7.	Appendices	79
7.1	Individual Results – Local Processing	79
7.2	Individual Results – Global Processing	85
7.3	Individual results – Combined Local and Global Processing	90

List of Figures

1.1	An Intermediate Representation in a Speech Processing System . . .	12
1.2	An Overview of the Processing	14
1.3	Regions of Influence of Phonemes	18
2.1	Broad Classes In Terms of Primary Features	25
3.1	A Schematic Representation of the Local Processing	29
3.2	Broad Class Model Topology	31
3.3	Topology for an Individual Class Within the Broad Class Model . . .	32
3.4	Two-alternative, Forced-choice Decision Model	39
6.1	Spectrogram with Truth Label Superimposed	76
6.2	Block Diagram for the Incorporation of Feedback	78

List of Tables

- 1.1 Feature Matrix Representations of “did you” 12
- 2.1 IPA Equivalents of TIMIT Labels 22
- 2.2 Linguistic Features of Vowels 23
- 2.3 Linguistic Features of Glides, Liquids, Nasals, and Affricates 24
- 2.4 Linguistic Features of Plosives and Fricatives 24
- 3.1 Mapping from TIMIT Label to Broad Class Label 30
- 3.2 Confusions Among Broad Class Estimates 34
- 3.3 Probability of a Correct Response vs. *d'* 41
- 3.4 Performance for Individual Features Using Local Processing 42
- 4.1 Performance for Individual Features in a Variety of Paradigms 52
- 4.2 Percent Change in Performance Between Paradigms 53
- 4.3 Agreement of Local and Global Processing 54
- 4.4 Reliability of Local Estimate When Global Estimate Disagrees 54
- 5.1 Confusion Matrix in Phoneme Classification 65
- 5.2 Confusion Matrix in Phoneme Classification 66
- 5.3 Confusion Matrix in Phoneme Classification 67
- 6.1 Assignment of Non-linguistically-motivated Features 73
- 6.2 Assignment of Non-linguistically-motivated Features 73
- 6.3 Assignment of Non-linguistically-motivated Features 74
- 7.1 Performance of Dependent Local Processing - Known Boundaries 80
- 7.2 Performance of Independent Local Processing - Known Boundaries 81
- 7.3 Performance of Local Processing - Known Boundaries 82
- 7.4 Performance of Local Processing - Unknown Boundaries 83

7.5	Performance of Local Processing - Unknown Boundaries	84
7.6	Performance of Global Processing - Known Boundaries	86
7.7	Performance of Global Processing - Known Boundaries	87
7.8	Performance of Global Processing - Boundaries Unknown	88
7.9	Performance of Global Processing - Boundaries Unknown	89
7.10	Performance of Local and Global Processing - Known Boundaries . .	91
7.11	Performance of Local and Global Processing - Known Boundaries . .	92
7.12	Performance of Local and Global Processing - Boundaries Unknown .	93
7.13	Performance of Local and Global Processing - Boundaries Unknown .	94

1. Introduction

1.1 The Goal and Its Motivation

Linguists describe a phoneme as a shorthand notation for a set of features which describe the operations of the articulators required to produce the meaningful aspects of a speech sound. In this thesis we develop a method of representing the speech waveform in terms of the same set of distinctive linguistic features, rendering it appropriate for a linguistically-motivated method of lexical access in the task of continuous speech recognition.

Most research in the area of automatic speech recognition has bypassed the representation question. That is, complex systems have been devised which are appropriate for the modeling of a large class of dynamic processes, but the fact that speech has linguistic structure has been largely ignored. The work described in this thesis, on the other hand, incorporates knowledge of the structure of the speech signal as well as a linguistic feature representation of the abstract sounds of speech to provide a representation of the speech waveform in terms of linguistic features.

Recognition is simply a representation at a certain level of abstraction. For example, a hidden-Markov-model-based continuous speech recognition system (HMM) with a null grammar finds the most likely sequence of lexical items to represent a waveform, thereby representing the original signal on the word level. With a language model, an HMM represents the waveform at the phrase level. Referring to figure 1.1, we develop an intermediate interface (data abstraction layer) between the physical (waveform) and application layers of speech processing, thereby adopting a conceptually different view of the task of recognition. While this thesis is concerned

with the transformation from the waveform to the feature representation, the motivation for the transformation comes from comparing the interfaces with the application layer from the physical and intermediate layers.

Current automatic speech recognition systems, which correspond to the left path in the figure, represent lexical entries in terms of a phonemic spelling and access words in terms of sequences of phonemes. This representation, however, disregards some of the phenomena which occur in conversational speech. In particular, relaxation of requirements on the production of a particular feature may occur. The following discussion is patterned after one given by Stevens [26]. Consider the expression “did you” which, when pronounced carefully, corresponds to the phonemes [D-IH-D-Y-UW]. When pronounced casually, however, the result may correspond to the phonemes [D-IH-JH-UH]. Phonemically, a considerable change has taken place in going from the theoretical representation of the expression and the representation corresponding to the utterance produced. Table 1.1 provides a representation of each of the pronunciations in terms of linguistic features, as will be described in Chapter 2. In the feature representation of the utterances, we see that the matrix entries remain largely intact in going from the first pronunciation to the second, with only the features anterior and strident changing in the collapsing of the D-Y to JH and the feature tense changing in the final vowel. The task of recovering the word sequence is more tractable from the second representation than from the first.

Related to the method of lexical access enabled by any representation is the notion of distance between phonemes implied by it. In the feature representation, distance reflects directly phonemic differences, while distance in the waveform space is taken as geometric distance between spectra which may be swamped with differences which are not phonemically relevant. For example, while one may feel that the phonemes “m” and “b” are close in some perceptual space, these sounds are quite different spectrally. In the feature representation, however, they differ in only one feature, so that the intuitive proximity is captured.

Furthermore, a feature representation of the speech waveform allows for a means of including rules of assimilation and transitional representations. Features do not

	D	IH	D	Y	UW	D	IH	JH	UH
VOCALIC	-	+	-	-	+	-	+	-	+
CONSONANTAL	+	-	+	-	-	+	-	+	-
HIGH	-	+	-	+	+	-	+	+	+
BACK	-	-	-	-	+	-	-	-	+
LOW	-	-	-	-	-	-	-	-	-
ANTERIOR	+	-	+	-	-	+	-	-	-
CORONAL	+	-	+	-	-	+	-	+	-
ROUND	-	-	-	-	+	-	-	-	+
TENSE	-	-	-	-	+	-	-	-	-
VOICE	+	+	+	+	+	+	+	+	+
CONTINUANT	-	+	-	+	+	-	+	-	+
NASAL	-	-	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	-	+	-
LABIAL	-	-	-	-	-	-	-	-	-

Table 1.1: Feature matrices for careful as well as casual pronunciations of “did you.”

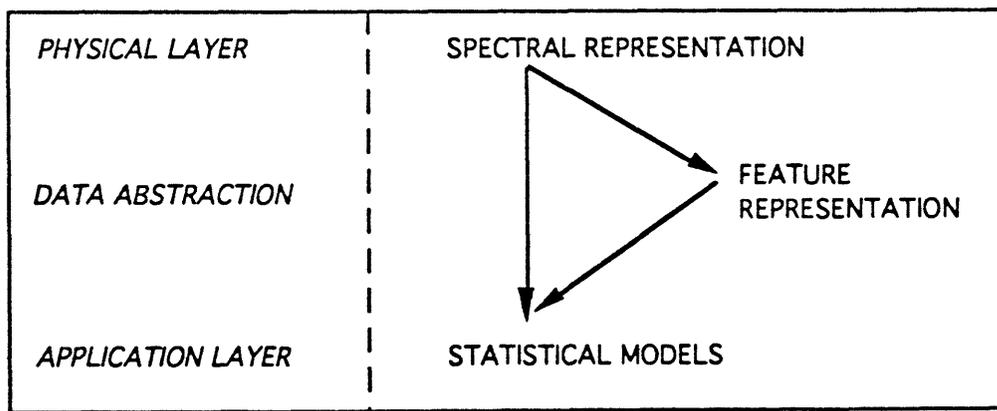


Figure 1.1: The introduction of an intermediate representation in a speech processing system.

change simultaneously in transitions from one sound to the next, but spread in a predictable manner through a transitional region. Therefore, it is possible to describe the waveform near transitions in terms of the feature configurations expected, rather than restricting the representations to be consistent with theoretical phoneme configurations. In fact, Deng [7] has designed an HMM structure in which states correspond to sets of feature configurations; states associated with transitional regions explicitly capture the spread of features.

Finally, the linguistic feature representation of the waveform is low-dimensional, reflecting directly the state of the speaker's articulatory system as a function of time.

The goal of this thesis, then, is to devise a means of parameterizing the speech waveform in terms of linguistic features which relies upon physical, statistical, and linguistic considerations, to demonstrate the efficacy of the procedure through the applications of phoneme classification and phoneme recognition, and to review the results of the procedure in order to gain insight into the amount of information about individual features provided by different methods of analyzing the speech waveform.

The main contribution of the work is that it provides a representation of the speech waveform appropriate for lexical access on the basis of features in the task of continuous speech recognition.

1.2 Terminology

We adopt the following terminology throughout the thesis:

PHONEME An abstract speech unit representing a specific mode of the speaker's articulatory system.

(SPEECH) SOUND Equivalent to phoneme.

(LINGUISTIC) FEATURE A component of a phoneme representing a state variable in the speech production system.

BROAD CLASS A set of phonemes which share the primary features listed in Chapter 2.

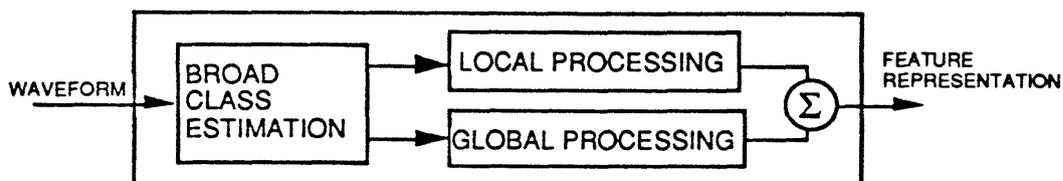


Figure 1.2: A schematic overview of the processing performed in order to assess the probability of each linguistic feature being represented in the neighborhood of each time frame.

PHONE A portion of the acoustic waveform in which all frames receive their principal influence from the same phoneme and neighboring frames receive influence principally from a different phoneme. Thus, by our definition, phones do not overlap in time although the manifestations of phonemes do. The region of overlap, however, may be different for different features.

SEGMENT Equivalent to phone.

FRAME One 5ms sample of a phone.

1.3 Structure of the Thesis

The remainder of this chapter describes the TIMIT database and provides an interpretation of its labels.

Chapter 2 describes the linguistic representation of speech in terms of production features and provides a description of broad classes of speech sounds in terms of a subset of these features.

Chapters 3 and 4 describe the algorithm by which we represent the waveform in terms of linguistic features. A schematic overview of the procedure is provided in figure 1.2. The initial stage of the hierarchical processing estimates the broad class of speech sounds represented. Based upon this estimate, we make both local and global inquiries as to the nature of the feature composition in the neighborhood of each frame. The terms local and global are chosen to emphasize that probabilities of features for a given frame are derived from narrow as well as wider windows in time around that frame. The outputs of the two levels of processing are averaged in order

to arrive at the final estimate of the probability of each feature being encoded in the neighborhood of each frame.

Chapter 3 describes the broad class estimation stage of processing, as well as the temporally-local processing scheme by which we assign probabilities of features being encoded in the waveform in the neighborhood of each frame. The likelihood of each linguistic feature being encoded in the waveform at a given time is evaluated using broad-class-specific models.

Chapter 4 describes the temporally-global stage of processing. This consists of a transitional modeling algorithm which is top-down in the sense that feature probabilities are derived from a mapping of phoneme probabilities. Transitions are defined as points in time at which the estimated broad class changes. Explicit modeling of the transitional regions takes into account the information about the features represented in one region carried by frames outside of the TIMIT phone boundaries.

Implementation of a continuous speech recognition system which is consistent with our representation of speech is outside of the scope of our project. Therefore we test the fidelity of our representation on the intermediate tasks of phoneme identification and phoneme recognition, as described in Chapter 5.

In the final chapter the results of estimating the presence or absence of each linguistic feature are analyzed and enhancements to the algorithm are suggested based upon this analysis. A summary of the main contributions of the thesis is then provided.

1.4 Experimental Data

Experimentation is done using utterances from the Prototype Version (Training Set) of the TIMIT database [30], which provides five phonetically compact sentences (“sx”) and three natural phonetic sentences (“si”) from each of 290 speakers across eight dialects.

1.4.1 Data Partitioning

In order to compare the performance resulting from different methods of assessing feature probabilities, we designated a subset of the database as a development set. For estimating model parameters, a portion of this set was defined as the development training set. The remainder of the set was used for evaluation of various algorithms. The development training set consisted of the sentences from the first third of the speakers (alphabetically) from each of the eight dialects, while the development testing set consisted of the sentences from the middle third of speakers within each dialect.

After deciding upon the final version of the algorithms, a final training and testing set were defined. The final testing set consisted of the sentences from the last two speakers in each dialect; the final training set consisted of all other sentences from the set of TIMIT male speakers. This training set, comprised of 2194 utterances, encompassed all of the development set. The results reported in the thesis derive from experiments on the final testing set of 128 utterances.

1.4.2 Waveform Representation

Several initial representations of the speech waveform were considered, including a set of attributes based on energy in various spectral bands and the time derivatives. Also considered were a set of cepstra and their time derivatives, a mixture of the two aforementioned representations, and a set of normalized cepstra and their time derivatives. The cepstral attributes are motivated by the source-filter view of speech production and tend to separate vocal tract properties from source characteristics [21]. As the normalized cepstra and derivatives resulted in the greatest accuracy in detecting the presence or absence of linguistic features, only the results from that representation will be discussed further.

The bandwidth of each utterance is 8kHz. Each time frame is parameterized by 14 normalized cepstra $NC_0 - NC_{13}$ and their time derivatives $DC_0 - DC_{13}$. In order to compute the normalized cepstra, the short-time Fourier transform was calculated

using a 5ms frame rate and a 15ms Hamming window. This window duration was chosen as a good compromise between the temporal resolution enabled by a short window and the frequency resolution provided by a long analysis window [21]. The squared magnitude of the transform was then mel-warped to provide greater resolution at low frequencies than at high, as is the case in the human auditory system [19]. LPC analysis was performed using 14 poles, providing smooth approximations of the warped spectra. The inverse Fourier transform of the logarithm of the result then provided the cepstral coefficients. Finally, the time derivative of each cepstral coefficient was approximated as the slope of the linear regression fit to the values of that coefficient over a 50ms window centered at the frame of interest. The time derivatives of the cepstra are included as a means of modeling consistent time variations of feature manifestations. These attributes have been shown to increase performance in other speech recognition tasks and seem less sensitive to variations between speakers than the cepstra themselves [11].

Normalization of the cepstra consisted of finding the median of each cepstral coefficient within the 80% highest energy frames of an utterance and subtracting that quantity from the coefficient calculated at each time frame. The lowest 20% energy frames represent for the most part silence regions, which provide little normalization benefit [11]. We interpret the subtraction of the normalization term from each frame as the canceling of channel differences from speaker to speaker. Indeed, in the case of a stationary channel response convolved with the input waveform, the result in the frequency domain is a multiplication of the channel spectrum with the input spectrum. The logarithmic operation in computing the cepstrum turns that multiplication into an additive operation which is canceled through subtraction. Here we are assuming that every utterance is long enough so that we have similar spectra in each utterance and long-term differences between utterances are due only to channel variations.

1.4.3 Interpretation of Labels

Phonemes are discrete units which are arranged sequentially to form words. The physical manifestation of the production of these phonemes, however, results in an

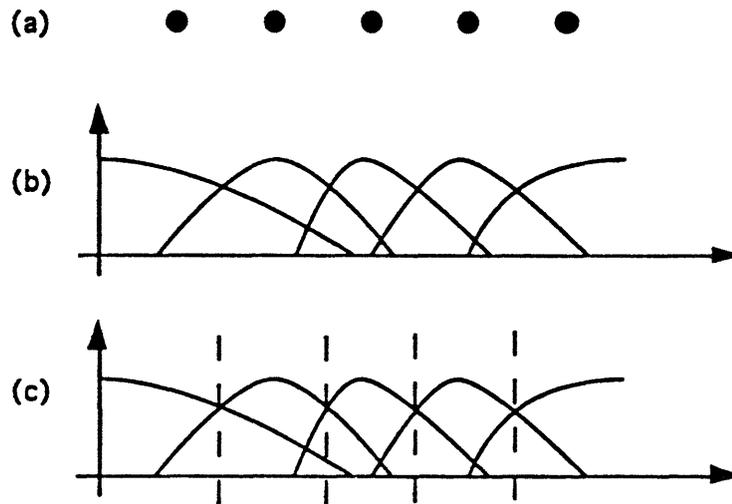


Figure 1.3: Schematic diagram of regions of influence of phonemes on the speech waveform.

overlap. Each phoneme has a region of influence in the waveform, so that the regions of influence of all of the phonemes cover, rather than partition, the waveform in time. This phenomenon is depicted schematically in figure 1.3. In panel (a) we show a sequence of dots, representing the arrangement of phonemes to form a word. In panel (b) we show the distribution of influence on the waveform over time for each of the phonemes. The regions where two or more phonemes exert an appreciable influence on the waveform are what we refer to as transitional. In panel (c) we impose dashed lines to indicate the regions in the waveform in which each of the phonemes exerts more of an influence than any of the other phonemes. This segmentation is assumed to be consistent with that given by the TIMIT markings. The resulting labels do not reflect the fact that feature spread from one phone to another will cause the spectra to exhibit some properties associated with the neighboring sound. We make use of the information about a sound present in the waveform outside of the TIMIT boundaries by explicitly modeling transitional regions, as will be described in Chapter 4.

2. A Distinctive Feature Representation of Phonemes

2.1 Introduction

Spoken words are composed of phonemes in the same manner that written words are composed of letters; as handwritten script bears the characteristics of an individual writer, acoustic realizations of phonemes bear characteristics specific to an individual speaker. For the sake of minimal effort in generation, both handwritten text and continuous speech trains are subjected to a deformation of the individual building blocks of the message in order to smoothly link components into a unified chain. However, in both handwriting and speech, the variations to the prototypical building blocks must not be so large as to distort the inherent qualities of the units if the result is to be understandable by other individuals. This fact suggests a description of the characters in terms of a set of attributes which are preserved under allowable deformations of the generic unit.

Just as letters may be described in terms of the strokes of the pen needed to produce them or by the manifestations of these actions such as line segments and curves, phonemes may be described in terms of the actions of a speaker needed to produce them, as well as time-varying frequency spectra which result from these actions. Indeed, in this thesis, the second representation is viewed as a set of observations which provide information about the first; estimates of the speaker-independent actions required to produce a sound are derived from the speaker-dependent acoustic manifestations.

2.2 A Linguistic Feature Set

The set of features which distinguish English phonemes is not unique; several sets have been introduced in the literature. The set which we shall adopt is, for the most part, that of Chomsky and Halle [4]. Specifically, we consider the following linguistic features, defined in terms of the required actions of a speaker in producing that sound and accompanied by specific spectral characteristics:

VOCALIC Sounds produced with an unstricted oral cavity and with vocal cords which are positioned so as to allow spontaneous voicing. Vocalic sounds are typically loud in relation to non-vocalic sounds and exhibit visible formants.

CONSONANTAL Includes sounds produced by forming an obstruction in the midsagittal region of the vocal tract, resulting in a lower total energy and lower first formant than non-consonantal sounds.

HIGH Sounds produced with the tongue body near the palate, resulting in a lowered first formant.

LOW Sounds produced with the tongue and jaw lowered, resulting in a high first formant.

BACK Includes those sounds produced with the tongue body toward the back of the mouth, resulting in a lowered second formant.

ANTERIOR Sounds produced with a constriction of the vocal tract anterior to the alveolar ridge.

CORONAL Includes those sounds for which the tongue blade is raised.

ROUND Sounds produced with rounded lips, causing all formants to lower in frequency.

TENSE Sounds produced with a deliberate and accurate gesture. Tense sounds are typically longer in duration with more extreme formant positions than non-tense sounds.

VOICE Sounds produced with the vocal folds vibrating, causing spectral resonances to become visible.

CONTINUANT Includes sounds for which the primary constriction of the vocal tract is not so narrow as to block the air flow past it, resulting in a smooth transition between the spectra associated with its predecessor and the spectra representing a continuant sound.

NASAL Sounds produced with the velum open. For nasal consonants in murmur the second formant is low in intensity and formant bandwidths are wide. Nasalized vowels typically exhibit an additional resonance below the first vowel formant, and simultaneous weakening and shift up in frequency of that formant.

STRIDENT The air stream is directed against an obstructing surface, resulting in a noisy spectrum with substantial high-frequency energy.

LABIAL The primary constriction is formed at the lips, leading to a lowered first and second formant.

2.3 Feature Configurations of Phonemes

We refer to phonemes by the typewritten symbols used for labeling the TIMIT database; the equivalent IPA symbols are given in table 2.1. The underlined letters in the words in the table provide sample occurrences of each of the phonemes.

Tables 2.2 through 2.4 depict the binary linguistic feature representation of each of the vowels and the consonants distinguished in this thesis. The “+” and “-” entries in the tables indicate the state of the corresponding articulator in the production of the sound. For example, sounds which are formed by rounding the lips are “+ round” while sounds which do not involve lip rounding are “- round.” Note that diphthongs have been excluded from the set of phonemes, as they consist of a transition from one feature vector to another, and therefore are the concatenation of two phonemes in this representation. In addition, neutral vowels have been omitted as the feature

TIMIT	IPA	EXAMPLE	TIMIT	IPA	EXAMPLE	TIMIT	IPA	EXAMPLE
ix	ɪ	re <u>ason</u>	iy	i ^y	be <u>t</u>	dx	ˈr	au <u>dit</u>
uw	U	sh <u>ampoo</u>	ux	ü	ne <u>w</u>	j	ʃ̣	<u>J</u> une
ey	e ^y	st <u>ay</u>	ow	o ^w	bo <u>at</u>	f	f	<u>f</u> ee
aa	ɑ	h <u>ot</u>	ih	ɪ	h <u>i</u> t	sh	ʃ̣	<u>s</u> he
uh	U	h <u>oo</u> d	eh	ɛ	be <u>t</u>	dh	ð	<u>th</u> ese
ah	ʌ	r <u>ug</u>	ao	ɔ	bo <u>u</u> ght	zh	ʒ̣	gar <u>ag</u> e
ae	æ	h <u>a</u> t	ax	ə	o <u>f</u>	kcl	k [□]	<i>k-closure</i>
ay	a ^y	m <u>y</u>	ax-h	ə̣	de <u>st</u> roy	tcl	t [□]	<i>t-closure</i>
aw	o ^w	h <u>ow</u>	oy	ɔ ^y	to <u>y</u>	bcl	b [□]	<i>b-closure</i>
er	ɜ̣	sh <u>ir</u> t	axr	ɝ̣	o <u>ff</u> er	g	g	g <u>o</u>
y	y	ye <u>llow</u>	w	w	wa <u>y</u>	ch	č̣	<u>ch</u> ew
l	l	l <u>e</u> ss	el	ɪ̣	ex <u>a</u> mple	s	s	s <u>e</u> e
r	r	r <u>e</u> d	hv	ɦ̣	h <u>a</u> ve	th	θ̣	w <u>it</u> h
hh	h	h <u>ow</u>	n	n	pl <u>a</u> n	z	z	z <u>oo</u>
en	ɲ̣	B <u>ost</u> on	nx	ɹ̣	mo <u>n</u> ey	v	v	v <u>e</u> ry
em	ṃ	co <u>m</u> plete	m	m	m <u>e</u>	pcl	p [□]	<i>p-closure</i>
ng	ŋ̣	si <u>ng</u>	k	k	ke <u>y</u>	gcl	g [□]	<i>g-closure</i>
t	t	to <u>o</u>	p	p	pa <u>y</u>	dcl	d [□]	<i>d-closure</i>
d	d	da <u>y</u>	b	b	be <u> </u>	q	ʔ	<i>glottal stop</i>

Table 2.1: The TIMIT label, equivalent International Phonetic Alphabet symbol, and a sample word for each of the phonemes discussed in this thesis.

configurations for these sounds are volatile and “H” has been excluded because of its difficulty in fitting into the linguistic feature framework [9]. We include a representation of quiet in order to represent with closures in the same framework as the phonemes. We follow the TIMIT notation of treating stop gaps as separate entities from the release even though linguistically these two units together comprise a single phoneme.

	VOWELS										
	IY	UW	EY	OW	AA	IH	UH	EH	AH	AO	AE
VOCALIC	+	+	+	+	+	+	+	+	+	+	+
CONSONANTAL	-	-	-	-	-	-	-	-	-	-	-
HIGH	+	+	-	-	-	+	+	-	-	-	-
BACK	-	+	-	+	+	-	+	-	+	+	-
LOW	-	-	-	+	+	-	-	-	-	-	+
ANTERIOR	-	-	-	-	-	-	-	-	-	-	-
CORONAL	-	-	-	-	-	-	-	-	-	-	-
ROUND	-	+	-	+	-	-	+	-	-	+	-
TENSE	+	+	+	+	+	-	-	-	-	-	-
VOICE	+	+	+	+	+	+	+	+	+	+	+
CONTINUANT	+	+	+	+	+	+	+	+	+	+	+
NASAL	-	-	-	-	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	-	-	-	-	-
LABIAL	-	-	-	-	-	-	-	-	-	-	-

Table 2.2: Linguistic features for each of the vowel sounds considered in this thesis.

Modern linguistic theory has departed from the notion of each phoneme being represented by the entire set of features. For example, since the production of vowels does not involve blocking the air flow through the vocal tract, the use of the feature continuant to describe vowels is unnecessary. The reduction of the representation to the non-redundant features describing each phoneme is efficient for the purposes of coding. However, from the viewpoint of recognition, the redundancies are desirable for recovery from errors as well as algorithm simplicity. We include the full set of feature descriptors for each phoneme as a sort of place keeper which will allow mathematical manipulation of our results, in much the same way that vectors lying in the x-y plane are specified as $[x, y, 0]$ in three dimensions.

Furthermore, some linguists now shun the notion of “feature bundles,” which

	GLIDES		LIQUIDS		NASALS			AFFRICATES		QUIET
	Y	W	L	R	M	N	NG	CH	JH	H#
VOCALIC	-	-	+	+	-	-	-	-	-	-
CONSONANTAL	-	-	+	+	+	+	+	+	+	+
HIGH	+	+	-	-	-	-	+	+	+	-
BACK	-	+	-	-	-	-	+	-	-	-
LOW	-	-	-	-	-	-	-	-	-	-
ANTERIOR	-	-	+	-	+	+	-	-	-	-
CORONAL	-	-	+	+	-	+	-	+	+	-
ROUND	-	+	-	-	-	-	-	-	-	-
TENSE	-	-	-	-	-	-	-	-	-	-
VOICE	+	+	+	+	+	+	+	-	+	-
CONTINUANT	+	+	+	+	-	-	-	-	-	+
NASAL	-	-	-	-	+	+	+	-	-	-
STRIDENT	-	-	-	-	-	-	-	+	+	-
LABIAL	-	-	-	-	+	-	-	-	-	-

Table 2.3: Linguistic features for each of the glide, liquid, nasal, and affricate phonemes considered, as well as the linguistic feature description of quiet.

	PLOSIVES							FRICATIVES						
	P	B	G	T	D	K	F	V	TH	DH	S	Z	SH	ZH
VOCALIC	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CONSONANTAL	+	+	+	+	+	+	+	+	+	+	+	+	+	+
HIGH	-	-	+	-	-	+	-	-	-	-	-	-	+	+
BACK	-	-	+	-	-	+	-	-	-	-	-	-	-	-
LOW	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ANTERIOR	+	+	-	+	+	-	+	+	+	+	+	+	-	-
CORONAL	-	-	-	+	+	-	-	-	+	+	+	+	+	+
ROUND	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TENSE	-	-	-	-	-	-	-	-	-	-	-	-	-	-
VOICE	-	+	+	-	+	-	-	+	-	+	-	+	-	+
CONTINUANT	-	-	-	-	-	-	+	+	+	+	+	+	+	+
NASAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	+	+	-	-	+	+	+	+
LABIAL	+	+	-	-	-	-	+	+	-	-	-	-	-	-

Table 2.4: Linguistic features for each of the plosive and fricative sounds considered.

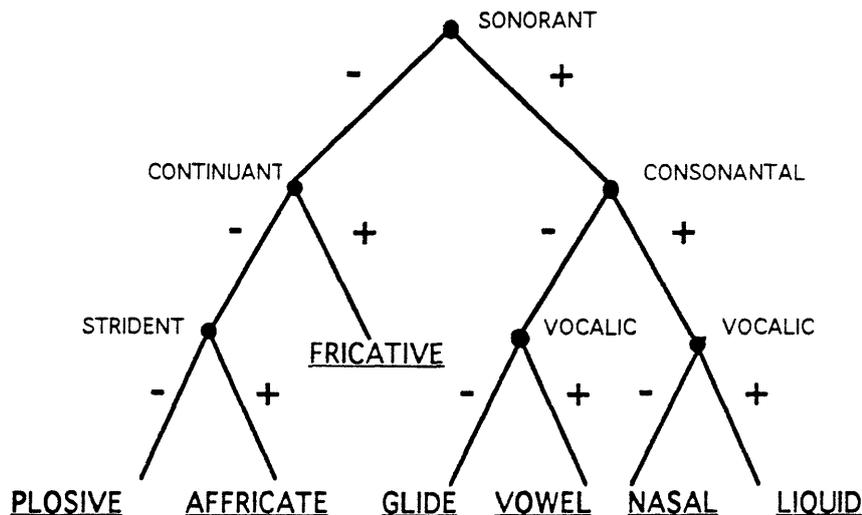


Figure 2.1: Broad classes as leaves of a tree built from primary features.

connotes a lack of structure of the features, in favor of the notion of feature geometries [22], [15]. Studies in geometries attempt to define hierarchical structures whose terminal nodes are the distinctive features. Nodes of the tree are defined so that rules of assimilation apply to children of that node but not across nodes.

From the viewpoint of recognition, the assimilation rules could prove useful in adjusting the training labels. That is, where we have used the one-to-one mapping of phonemic label to feature presence or absence assimilation rules could provide one-to-many mappings in that certain features are assimilated to take on the value associated with a neighboring sound. These rules are not encoded in this thesis. The problem of assimilation is somewhat, although not fully, addressed through the TIMIT labeling system in that the label assigned is that actually spoken, not that expected out of context. For example, if the nasal consonant in “can be” is labialized by the speaker, we expect the TIMIT label to be “M” rather than “N.” However, when the assimilation does not map a set of features onto another phoneme label, the affected feature change is missed.

For the purposes of recognition, we adopt the following model of feature organization. We include the features sonorant and instant release [9] which are redundant given the representation shown in tables 2.2-2.4, but which aid in the definition of broad classes in terms of features. The feature sonorant is characterized by a lack of

pressure built up in producing a sound; the feature instant release specifies whether built-up pressure is released immediately or through a more gradual release [1].

- *Primary features:* Sonorant, Vocalic, Consonantal, Instant Release, Continuant
- *Secondary features:* Strident, Tense, High, Back, Low, Anterior, Coronal, Labial, Voice, Round, Nasal

We contend that the primary features determine the gross spectral characteristics of the resulting speech waveform; the other features modulate or make fine structural changes to the basic pattern defined by the primary ones. Therefore sounds which are characterized by the same primary features have similar spectra qualitatively, while sounds which have different primary features are fundamentally different. This fact implies that the features are encoded in the waveform hierarchically, with the manifestation of secondary features dependent on the configuration of the primary ones. For example, because G and IY have different configurations of primary features, the feature +high will be encoded differently in the waveform for the two phonemes. Consider the structure shown in figure 2.1. The broad classes we distinguish along the tops of tables 2.2- 2.4 are the leaves of the structure, and the internal nodes branch according to the presence or absence of the primary features. The two-stage hierarchical search for features which is described in Chapter 3 is essentially a search first for the manifestations of encoding a set of primary features in the neighborhood of each frame and then, given those features, a search for the secondary features, as well as a verification of the primary ones. The estimation of the broad class as a whole, as will be described, is meant to capture dependencies among the primary features.

3. A Distinctive Feature Representation – Local Processing

3.1 Overview of the Procedure

Chapter 2 described phonemes in terms of a set of linguistic features; the procedure outlined in this chapter has as its aim assessing the probability that a frame of speech represents a phoneme in which a given linguistic feature is present using information gathered only from the neighborhood of that frame. The output of this procedure will be combined with a more global processing scheme which will be discussed in the next chapter. The terms “local” and “global” are chosen to emphasize that the latter takes into account contextual information whereas the former does not, as well as the fact that the two procedures involve processing at different time scales. Results discussed in this chapter are those in which the global processing stage is omitted, so that the probabilities estimated from the local stage are taken as the probabilities of the presence of the features.

The acoustic manifestation of a linguistic feature is dependent on the broad class of speech sounds to which the principally-represented phoneme belongs. For example, as mentioned in the previous chapter, the feature +high will be encoded in the speech waveform differently in the cases of the vowel IY and the plosive G. Furthermore, all frames of a phone which corresponds to the presence or absence of a feature need not be spectrally similar. To capture time variations of the acoustic correlates of the features, we model separately the beginning, middle, and ending portions of the phones representing each broad class. A schematic representation of the local

processing is given in figure 3.1. The oval region represents the space of all speech frames; pie-shaped wedges represent the segregation of those frames into the broad class most-likely represented. Time in the diagram progresses radially outward, so that the intersection of a ring with a wedge corresponds to the time portion (in thirds) of a broad class most likely represented. After each frame has mapped to a sector of the diagram, the probability of each feature is estimated from Gaussian models parameterized from training data which were also *estimated* to belong to that region. Thus, we have 14 feature present and 14 feature absent models within each of the 24 sections of the diagram. In order to assess the probability of a given feature, for example round, for a test frame estimated to represent the end of a vowel, we form the likelihood ratio from the models for +round and -round built from training samples also estimated to represent the final third of a vowel. We use the estimated broad class portion rather than the true label associated with each frame in the training set to provide a means of recovery from errors in the broad class estimation stage. For example, if models for +consonantal in the vowel regions were parameterized using training tokens which truly represented vowels, then the models for this feature would be empty, as no vowel is +consonantal. However, if we use the estimated broad class along with the true feature configuration, then we have the opportunity to model, for example, frames corresponding to an “L” or “R” which were wrongly estimated to represent vowels. Indeed, we have found experimentally that using the estimated broad class to segregate the training data led to a slightly higher accuracy in assessing feature probabilities than segregating the training set according to the true broad class labels.

3.2 Gaussian Models for Broad Class Estimation

As outlined in the previous section, each time frame t in the training set is assigned a truth label, $\tau(t) \in \{1, \dots, 8\}$, reflecting the broad class of speech sounds which that frame represents. The label is the result of a many-to-few mapping of the TIMIT labels, as indicated in table 3.1. Sounds grouped by parentheses represent an

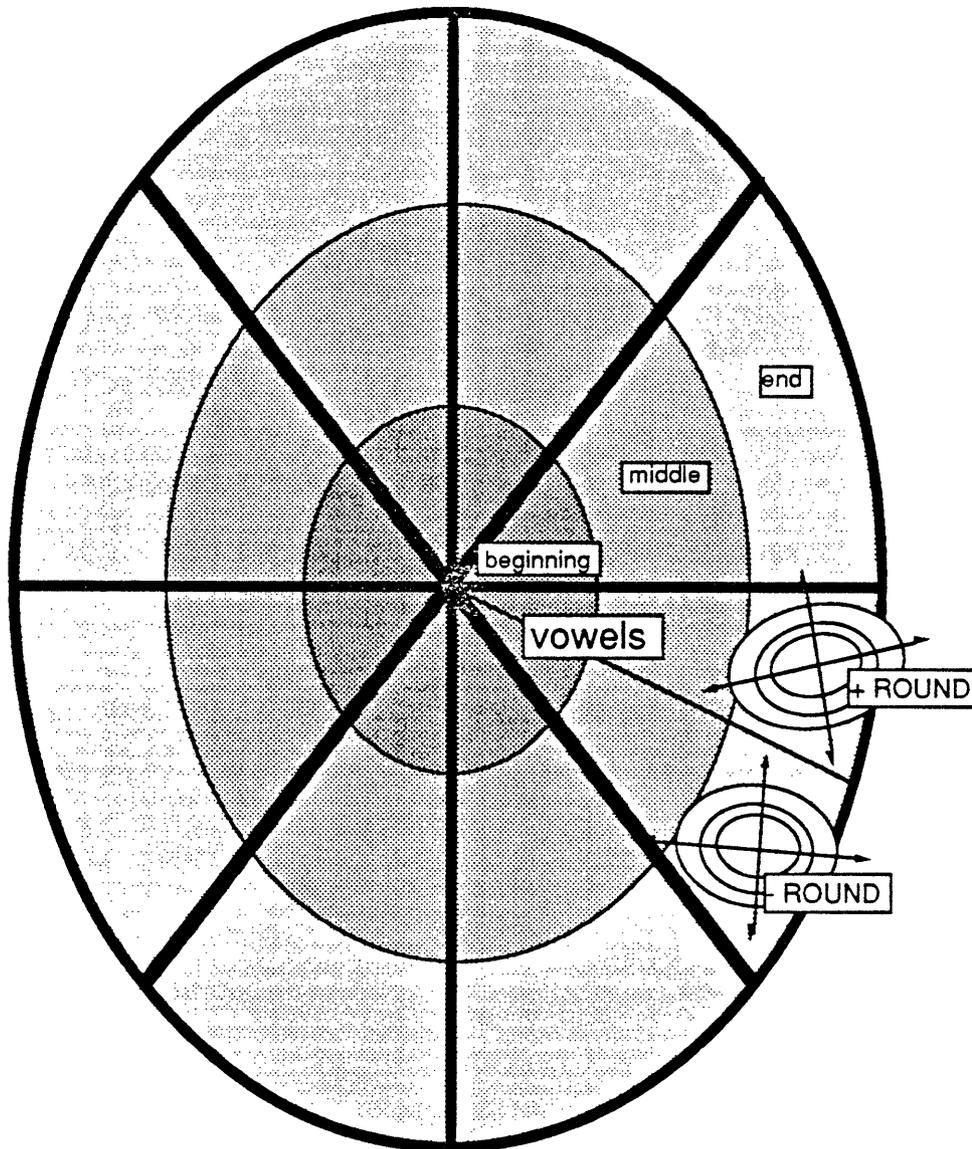


Figure 3.1: A schematic representation of the local stage of processing.

allophonic equivalence class which we treat by mapping all of the labels in the class to the first label of the group.

TYPE OF SOUND	CONSTITUENT TIMIT LABELS
VOWEL	IY, UW, EY, OW, AA, IH, UH, EH, AH, AO, AE
GLIDE	Y, W
LIQUID	(L, EL), (R, ER, AXR)
NASAL	(M, EM), (N, EN), NG
PLOSIVE	P, B, G, T, D, K
AFFRICATE	CH, JH
FRICATIVE	F, V, TH, DH, S, Z, SH, ZH
QUIET/VOICE BAR	(H#, PCL, TCL, KCL, BCL, GCL, DCL, EPI, PAU)

Table 3.1: Mapping from TIMIT label to broad class label.

Furthermore, each frame t of the training set is assigned a section label, $\sigma(t) \in \{1, 2, 3\}$, indicating whether the frame represents the initial, middle, or final third of a phone. Each phone is divided into three pieces of equal duration in order to enable modeling of the time variation of the manifestation of a feature within each broad class.

The broad class and section labels are used to segregate the frames in the training set for the purposes of parameterizing Gaussian models. For section $i \in \{1, 2, 3\}$ of phones representing broad class $k \in \{1, \dots, 8\}$ we estimate the model parameters as follows:

$$\begin{aligned}
 N_{k_i} &= \sum_{\{t|\sigma(t)=i, \tau(t)=k\}} 1 \\
 \hat{\mu}_{k_i} &= \frac{1}{N_{k_i}} \sum_{\{t|\sigma(t)=i, \tau(t)=k\}} x(t) \\
 \hat{\Sigma}_{k_i} &= \frac{1}{N_{k_i}} \sum_{\{t|\sigma(t)=i, \tau(t)=k\}} x(t) x^T(t) - \hat{\mu}_{k_i} \hat{\mu}_{k_i}^T
 \end{aligned}$$

That is, N_{k_i} is the total number of frames in the training set estimated as being drawn from the i -th third of a phone representing a phoneme in broad class k , $\hat{\mu}_{k_i}$ is the sample average (vector) over those same frames, and $\hat{\Sigma}_{k_i}$ is the sample covariance of that set of frames. We have that $N_{k_1} \approx N_{k_2} \approx N_{k_3}$ with differences arising due only

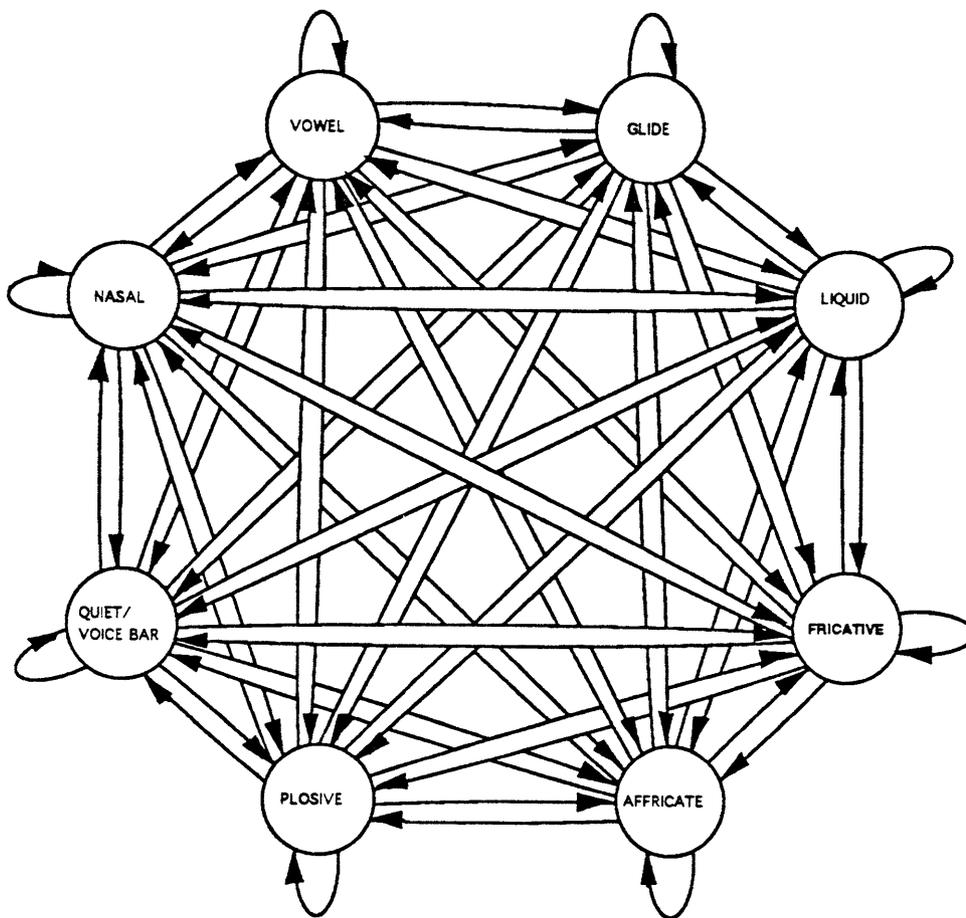


Figure 3.2: Broad class model topology.

to roundoff errors in dividing phones into thirds.

Given that section i of a phone representing broad class k is being produced, the N -dimensional probability density function for observations is taken as:

$$p(x|\tau = k, \sigma = i) = \frac{1}{(2\pi)^{\frac{N}{2}} |\hat{\Sigma}_{k,i}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\hat{\mu}_{k,i})^T \hat{\Sigma}_{k,i}^{-1} (x-\hat{\mu}_{k,i})} \quad (3.1)$$

3.3 Estimating Broad Classes

In order to estimate the sequence of portions of phones and the broad class which each phone represents in a particular sentence, the probabilities of the observed attribute vectors given each possibility are evaluated using equation 3.1 above.

A Markov model of broad class representation is used, as shown in figure 3.2. Each

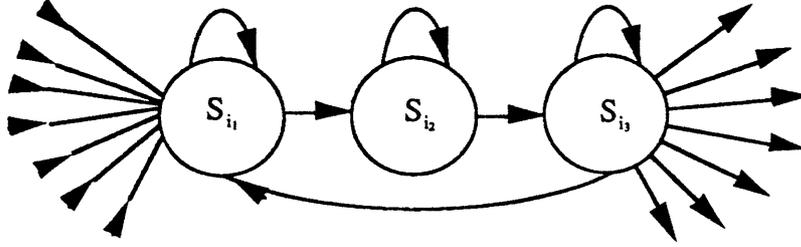


Figure 3.3: Topology for an individual class within the broad class model.

state in the diagram is actually composed of three states connected in a left-to-right fashion as in figure 3.3. These substates model explicitly the beginning, middle, and end of phones representing each broad class. To estimate the transition probabilities among states, the number of transitions on a frame-by-frame basis among the broad class and section labels from the training data are counted. If we define the number of transitions from state s_i to state t_j in the training set as $T_{s_i t_j}$, then the transition probability between these two states is estimated as the number of transitions from state s_i to state t_j divided by the total number of transitions from state s_i :

$$\hat{\alpha}_{s_i t_j} = \frac{T_{s_i t_j}}{\sum_{t=1}^8 \sum_{j=1}^3 T_{s_i t_j}}$$

Because each TIMIT sentence begins with silence, we assign initial state probabilities as:

$$\pi_{s_i} = \begin{cases} 1 & s = 8 \text{ and } i = 1 \\ 0 & \text{otherwise} \end{cases}$$

Finally, dynamic programming is used to find the most likely sequence of broad classes arising in each sentence. In the absence of known segment boundaries, for an utterance of length $M + 1$, we take the estimated sequence of states to be the most-likely trajectory through the Markov model:

$$S_0^*, \dots, S_M^* = \arg \max_{S_0, \dots, S_M} \pi_{S_0} \prod_{m=1}^M p(x(t) | S_m) \hat{\alpha}_{S_{m-1} S_m}$$

where $S_m \in \{s_i : s \in \{1, \dots, 8\}, i \in \{1, 2, 3\}\} \forall m$.

When phonemic boundary information is provided, as is the case for the task

of phoneme classification, the dynamic programming cost function is modified to reflect the fact that only one broad class is represented by the set of frames between each boundary set and that the region to the right of a boundary must represent the beginning of a broad class while the region to the left of a boundary represents the end of a broad class. The estimated sequence of broad class states through an utterance in the case of known boundary information is taken as:

$$S_0^*, \dots, S_M^* = \arg \max_{S_0, \dots, S_M} \pi_{S_0} \prod_{m=1}^M C(S_m | B_\alpha, B_\omega) p(x(t) | S_m) \hat{a}_{S_{m-1} S_m}$$

where the additional term $C(S_m | B_\alpha, B_\omega)$ reflects the cost of being in state S_m at time m given the set of boundary points marking the beginnings of phonemes, B_α , and the set of points marking the endings of phonemes, B_ω . Specifically,

$$C(S_m | B_\alpha, B_\omega) = \begin{cases} 0 & m \in B_\alpha, S_m \in \{s_i : s \in \{1, \dots, 8\}, i \in \{2, 3\}\} \\ 0 & m \in B_\omega, S_m \in \{s_i : s \in \{1, \dots, 8\}, i \in \{1, 2\}\} \\ 0 & m \notin B_\alpha \cup B_\omega, S_{m-1} \in \{s_i : s \in \{1, \dots, 8\}, i \in \{1, 2, 3\}\}, \\ & S_m \in \{t_j : t \in \{1, \dots, 8\}, j \in \{1, 2, 3\}\}, s \neq t \\ 1 & \text{otherwise} \end{cases}$$

We use the TIMIT boundary information for assigning broad class estimates to each training utterance, regardless of the availability of this information for the test set.

3.4 Broad Class Accuracy

In the case of known boundaries, we obtain the confusion matrix of broad classes shown in table 3.2, representing 84% correct classification of the frames in our final TIMIT test set. The broad classes listed vertically represent the true labels associated with each frame in the test set, while the classes listed across the top of the matrix represent the estimated class assigned to each frame. Thus, diagonal entries in the matrix correspond to correct estimates while the i, j -th entry represents an estimated

broad class of j for a frame taken from the true class i . As shown, a large proportion of the errors are among vowels, liquids and glides. If we were to collapse the categories vowel, glide, and liquid into one category such as vowel-like the broad class accuracy would be 90%. However, we choose to maintain these as separate broad classes on the basis of linguistic considerations as well as the fact that the larger number of classes allows for more flexible modeling of the linguistic features. The numbers indicated in the table were derived by collapsing the labels representing the thirds of each broad class into a single bin for that class.

	Fricative	Nasal	Vowel	Glide	Liquid	Affricate	Plosive	Quiet
Fricative	0.755	0.038	0.005	0.004	0.010	0.089	0.041	0.057
Nasal	0.007	0.889	0.022	0.024	0.018	0.000	0.011	0.028
Vowel	0.001	0.019	0.843	0.057	0.061	0.001	0.002	0.015
Glide	0.000	0.019	0.038	0.855	0.064	0.000	0.018	0.005
Liquid	0.000	0.033	0.125	0.101	0.711	0.001	0.009	0.020
Affricate	0.039	0.012	0.000	0.000	0.000	0.929	0.014	0.006
Plosive	0.029	0.004	0.000	0.007	0.003	0.097	0.817	0.043
Quiet	0.022	0.033	0.001	0.005	0.003	0.000	0.006	0.930

Table 3.2: Confusions among broad class estimates in the case of known boundary locations. Class presented is listed vertically; class estimated is listed horizontally.

3.5 Gaussian Models for Linguistic Features

The clustering information provided by the broad class estimation stage is used to build a set of estimated-broad-class-dependent feature vs. non-feature Gaussian models. Each frame $x(t)$ of the training set is assigned a label $\tau^f(t) \in \{0, 1\}$ indicating whether that frame corresponds to the absence ($\tau^f = 0$) or the presence ($\tau^f = 1$) of each linguistic feature $f \in \{1, \dots, 14\}$. All frames in the TIMIT training set which are *estimated* to be in a given portion of a phone representing a given broad class are divided into “feature present” and “feature absent” subgroups for each linguistic feature to be modeled.

Gaussian models of the waveform attribute vectors are parameterized for each

subgroup. We estimate the following model parameters:

$$\begin{aligned}
N_{k_i}^{f+} &= \sum_{\{t|S^*_t=k_i, \tau^f(t)=1\}} 1 \\
N_{k_i}^{f-} &= \sum_{\{t|S^*_t=k_i, \tau^f(t)=0\}} 1 \\
\hat{\mu}_{k_i}^{f+} &= \frac{1}{N_{k_i}^{f+}} \sum_{\{t|S^*_t=k_i, \tau^f(t)=1\}} x(t) \\
\hat{\mu}_{k_i}^{f-} &= \frac{1}{N_{k_i}^{f-}} \sum_{\{t|S^*_t=k_i, \tau^f(t)=0\}} x(t) \\
\hat{\Sigma}_{k_i}^{f+} &= \frac{1}{N_{k_i}^{f+}} \sum_{\{t|S^*_t=k_i, \tau^f(t)=1\}} x(t) x^T(t) - \hat{\mu}_{k_i}^{f+} \hat{\mu}_{k_i}^{f+ T} \\
\hat{\Sigma}_{k_i}^{f-} &= \frac{1}{N_{k_i}^{f-}} \sum_{\{t|S^*_t=k_i, \tau^f(t)=0\}} x(t) x^T(t) - \hat{\mu}_{k_i}^{f-} \hat{\mu}_{k_i}^{f- T}
\end{aligned}$$

where $N_{k_i}^{f+}$ indicates the number of frames which were estimated to belong to portion i of a phone representing broad class k and which represent the presence of feature f . $\hat{\mu}_{k_i}^{f+}$ and $\hat{\Sigma}_{k_i}^{f+}$ represent the sample mean and sample covariance, respectively, of this set of frames. Similarly, $N_{k_i}^{f-}$ indicates the number of frames which were estimated to belong to portion i of a phone representing broad class k and which represent the absence of feature f , with $\hat{\mu}_{k_i}^{f-}$ and $\hat{\Sigma}_{k_i}^{f-}$ representing the sample mean and sample covariance of these frames.

The N -dimensional probability density function of frames which represent the presence of a given feature for estimated portion i of a phone representing broad class k is modeled as:

$$p(x|k_i, \tau^f = 1) = \frac{1}{(2\pi)^{\frac{N}{2}} |\hat{\Sigma}_{k_i}^{f+}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\hat{\mu}_{k_i}^{f+})^T \hat{\Sigma}_{k_i}^{f+}{}^{-1} (x-\hat{\mu}_{k_i}^{f+})}$$

Similarly, the density function of frames representing the absence of a given feature

within the estimated broad class is modeled as:

$$p(x|k_i, \tau^f = 0) = \frac{1}{(2\pi)^{\frac{N}{2}} |\hat{\Sigma}_{k_i}^{f^-}|^{\frac{1}{2}}} e^{-\frac{1}{2}(x - \hat{\mu}_{k_i}^{f^-})^T \hat{\Sigma}_{k_i}^{f^-}^{-1} (x - \hat{\mu}_{k_i}^{f^-})}$$

3.6 Dependent Modeling of Place Features

The procedure described in the previous section treats each feature independently within each estimated broad class. In this section we describe a method for modeling jointly the features high, back, and low as well as anterior and coronal.

For training, we divide the set of frames in each estimated broad class according to an index \mathcal{I} defined as:

$$\mathcal{I} = \sum_{n=0}^{N-1} f_n 2^n$$

where N is the number of features in the group to be modeled dependently and f_n is the true value of the n -th feature of the phoneme underlying the frame. For example, in modeling the features anterior and coronal jointly, $N = 2$ and frames representing the phoneme M which is +anterior, -coronal map to an index $\mathcal{I} = 0 * 2^0 + 1 * 2^1 = 2$.

For each of the 2^N possible indices, we find the mean and covariance of the data mapping to that index and parameterize unimodal Gaussians accordingly. The relative number of frames in each set serves as a prior probability estimate.

Half of the total 2^N Gaussians then correspond to the presence of a given feature. We exploit this fact in assessing the probability that the feature is encoded. If we also define

$$\bar{\mathcal{I}}_m = \sum_{\{n \in \{0, \dots, N-1\} | n \neq m\}} f_n 2^n$$

for a given configuration of features in the group and \mathcal{R}_m as the set of values that $\bar{\mathcal{I}}_m$ takes on, then we have:

$$\begin{aligned} p(f_m = 1|x) &= \sum_{\{n \in \mathcal{R}_m\}} p(f_m = 1, \bar{\mathcal{I}}_m = n|x) \\ &= \frac{1}{p(x)} \sum_{\{n \in \mathcal{R}_m\}} p(x|f_m = 1, \bar{\mathcal{I}}_m = n) p(f_m = 1, \bar{\mathcal{I}}_m = n) \end{aligned}$$

Note that the feature labial is redundant for the combination +anterior, -coronal so that by modeling the dependencies of these features we automatically model the feature labial.

Note that the feature labial is redundant for the combination +anterior, -coronal so that by modeling the dependencies of these features we automatically model the feature labial. However, the model for -labial is different from that of the independent modeling under the dependent technique as it consists of a mixture of the models for the three other configurations of anterior and coronal.

3.7 Estimating Linguistic Features

For applications of the feature representation, vectors of probabilities of features are envisioned as the input. However, in order to assess the quality of the representation, we estimate the presence or absence of each linguistic feature for each time frame. The accuracy of the estimate derived from local processing, will be compared with that of the estimate derived from a combination of local and global processing in the next chapter.

For a given feature, each time frame in the speech waveform is associated with a probability of that linguistic feature being present using the Gaussian class vs. nonclass Gaussian models described above. The probabilities are associated with a two-state Markov model, where one state represents the encoding of a feature in the waveform and the second state represents the absence of the feature. The transition probabilities are functions of the broad class estimate.

To estimate the transition probabilities between the states, we count the number of each of the possible transitions on a frame-by-frame basis in the training data. If we define the number of transitions from state i to state j for estimated portion and broad class S^* as $T_{ij}^f(S^*)$ for $f \in \{1, \dots, 14\}$ and $i, j \in \{0, 1\}$, then the transition probability between these states is estimated as the number of transitions from state

i to state j for broad class S^* divided by the total number of transitions from state i :

$$\hat{\alpha}_{ij}^f(S^*) = \frac{T_{ij}^f(S^*)}{\sum_{j=0}^1 T_{ij}^f(S^*)}$$

Dynamic programming is used to find the most likely feature sequence arising in each sentence. For an utterance of length $M + 1$ with no segment boundary information given, and for each linguistic feature, we take the estimated sequence of corresponding feature presences to be:

$$P_{0}^{f*}, \dots, P_{M}^{f*} = \arg \max_{P_{0}^f, \dots, P_{M}^f} \pi_{P_{0}^f} \prod_{m=1}^M p(x(t) | P_{m-1}^f, S_m^*) \hat{\alpha}_{P_{m-1}^f P_m^f}^f(S_m^*)$$

where $P_m^f \in \{0, 1\} \forall m$. If segment boundary information is given, an additional cost term is included to reflect the fact that either all frames between the boundaries correspond to the presence of the feature, or all the frames between the boundaries correspond to the absence of that feature. Mathematically this is achieved by taking:

$$P_{0}^{f*}, \dots, P_{M}^{f*} = \arg \max_{P_{0}^f, \dots, P_{M}^f} \pi_{P_{0}^f} \prod_{m=1}^M C(P_m^f | B_\alpha, B_\omega) \cdot p(x(t) | P_{m-1}^f, S_m^*) \hat{\alpha}_{P_{m-1}^f P_m^f}^f(S_m^*)$$

where

$$C(P_m^f | B_\alpha, B_\omega) = \begin{cases} 0 & m \notin B_\alpha, P_m^f \neq P_{m-1}^f \\ 1 & \text{otherwise} \end{cases}$$

3.8 Results of Linguistic Feature Estimation

Shown in tables 7.1 through 7.3 of Appendix 1 is an indication of the algorithm's performance on an individual phoneme basis when phoneme boundary information is provided. For each of the phonemes listed in the left-hand column of the tables, the relative frequency of the frames corresponding to that phoneme which were estimated to represent the presence of the features listed horizontally is given. The + or - follow-

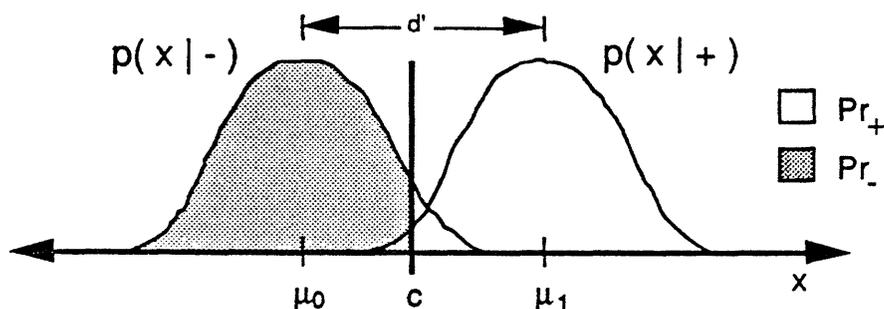


Figure 3.4: Two-alternative, forced-choice decision model.

ing each entry indicates the theoretical presence or absence of the feature. Table 7.1 indicates the results when the features high, back, and low as well as anterior and coronal are treated dependently as described in section 3.6, while table 7.2 shows the results when these features are modeled independently.

Tables 7.4 and 7.5 reflect the case of unknown phoneme boundaries. Dependent modeling of the features high-back-low and anterior-coronal was used to generate this result.

In this section we summarize the local processing performance, analyzing the results in terms of the separability of the models as described below.

3.8.1 Method of Analysis

For a quantitative analysis of our results, we have viewed feature estimates as if they arose from a psychophysical two-alternative forced-choice decision model. Referring to figure 3.4, we find the distance d' between the means of the two conditional density functions, $p(x|+)$ and $p(x|-)$, where x is some unobserved decision variable. Hypotheses H_+ and H_- correspond to the linguistic feature being encoded and not encoded in the neighborhood of a given frame, respectively. Unit variance is assumed for each model. This analysis, which is largely insensitive to bias, provides a measure of separability between feature present and feature absent models.

If we define Pr_+ to be the probability of estimating a given feature f to be present given that the feature is present and Pr_- to be the probability of estimating

the feature to be absent given that it is absent within a set of phonemes \mathcal{C} , we have:

$$N_+ = \sum_{\{\phi \in \mathcal{C} | \phi \ni f\}} Pr(\phi)$$

$$Pr_+ = \frac{1}{N_+} \sum_{\{\phi \in \mathcal{C} | \phi \ni f\}} Pr(+|\phi)Pr(\phi)$$

where $Pr(+|\phi)$ is the relative frequency of estimating frames representing chiefly phoneme ϕ to have the feature present and $Pr(\phi)$ is the prior probability of ϕ . We introduce \mathcal{C} so that we may restrict our attention to subsets of phonemes for a given feature; for example we may want to consider performance for the feature “tense” in vowels only.

Then:

$$Pr_+ = \int_c^\infty p(x|+) dx$$

$$= \int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_+)^2} dx$$

$$= \Phi(c - \mu_+)$$

where $\Phi(y) = \int_y^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$. Similarly,

$$Pr_- = \int_{-\infty}^c p(x|-) dx$$

$$= 1 - \int_c^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_-)^2} dx$$

$$= 1 - \Phi(c - \mu_-)$$

Thus we have that

$$\mu_+ = c - \Phi^{-1}(Pr_+)$$

$$\mu_- = c - \Phi^{-1}(1 - Pr_-)$$

and finally,

$$d' = |\Phi^{-1}(1 - Pr_-) - \Phi^{-1}(Pr_+)|$$

To get a feel for the meaning of various values of d' , we show in table 3.3 the probability of a correct response assuming equally-likely hypotheses for various values of d' .

d'	P_c	d'	P_c	d'	P_c
3.5	0.96	3.0	0.93	2.5	0.89
2.0	0.84	1.5	0.77	1.0	0.69

Table 3.3: Probability of a correct response for various values of d' .

3.8.2 d' Analysis of Local Processing

In table 3.4 we show the resulting d' for each feature under various experimental conditions. The results shown in the table take into account all of the phonemes modeled in calculating d' . Some discussions of d' within a subset of the phonemes are given in the analysis below. Condition A represents the case of local processing in which all features are modeled independently. Condition B corresponds to local processing with dependent modeling of the features high, back, and low as well as anterior and coronal. In this experiment entries not shown are identical to the corresponding value in column A. Column C indicates the results of local processing with unknown boundaries and dependent modeling of the place features.

Loss of boundary information caused the accuracy of detecting all features to decrease. This is to be expected in part because the results are presented on a frame by frame basis; if broad class estimates do not change at the TIMIT boundary points inappropriate models may be used to assess the probability of each feature for a few frames at the edge of each phone. Given below is a discussion of the results for individual features in which anomalies in performance were seen.

VOCALIC The feature vocalic was easily detected for most phonemes as evidenced

FEATURE	A	B	C
VOCALIC	3.807		3.010
STRIDENT	3.631		2.951
NASAL	3.503		2.910
CONSONANTAL	3.226		2.614
CONTINUANT	3.219		2.472
VOICE	2.758		2.440
ANTERIOR	2.705	2.666	2.119
ROUND	2.661		2.213
CORONAL	2.574	2.625	2.130
BACK	2.550	2.536	2.195
LOW	2.483	2.369	2.127
LABIAL	2.369		1.972
HIGH	2.278	2.442	2.038
TENSE	2.051		1.671

Table 3.4: Performance (d') for individual features using local processing. **A**: Independent processing, boundaries known. **B**: Dependent processing, boundaries known. **C**: Dependent processing, boundaries unknown.

by the relatively large value of d' in table 3.4. Notable exceptions were Y and W, whose vowel-like spectra caused them to be fairly frequently misclassified as +vocalic.

CONSONANTAL The main difficulty in assessing the value of the feature consonantal arose in the cases of UH, UW and L. For L, part of the problem might stem from the fact that frames labeled EL were mapped to the symbol L in determining feature labels; frames labeled EL in the TIMIT database act as vowels and do not involve a constriction of the vocal tract until the end of the sound. The vowels UH and UW were occasionally misclassified as liquids; the large bias toward +consonantal for frames in this estimated broad class led to an error in assessing the feature consonantal.

HIGH Dependent processing of the features high, back, and low with known boundaries led to a d' which was 7% larger than that achieved by independent processing for the feature high. On the subset of vowels, assuming each vowel to be equally likely, the dependent modeling of the feature high with low and back

led to a d' of 1.896, with independent modeling giving $d' = 1.499$. Difficulty in assessing the value of the feature high arose mainly for UW, IH, UH, EY, NG, K, and G. Again, as in the case of consonantal, difficulty for UW and UH stems from misclassification as liquids; since neither L or R is +high the bias toward -high is large within that estimated broad class. Difficulty for IH probably stems from its confusability with EH and AE, both of which are -high. Difficulty with the -high EY most likely stems from its spectral similarity to the +high IY, which is much more common. NG, K and G most likely suffer from being relatively uncommon, so that the biases in the models within nasals and plosives are largely toward -high.

BACK The feature back showed a 0.5% decrease in overall d' in going from independent to dependent local processing when boundary information was provided. Within the set of vowels, assuming each vowel to be equally likely, dependent modeling ($d' = 2.133$) and independent modeling (2.136) were comparable. The phonemes which were most difficult for which to assess the value of back were UW, UH, AH, L, NG, K, and G. Similarity to EH probably led to difficulty for AH while the problems discussed above surfaced again for the feature back in the case of the other anomalies.

LOW The feature low showed a 4.6% decrease in d' on the full set of phonemes in going from independent to dependent local processing when boundary information was provided. Within the subset of vowels, assuming each vowel to be equally likely, dependent processing ($d' = 1.772$) provided greater separability than independent processing ($d' = 1.571$). Difficulty in the feature low occurred mainly for OW and AA. Bias as well as spectral similarity to AO most likely caused the difficulties.

ANTERIOR For the feature anterior, a 1.4% decrease in overall d' occurred in going from independent local processing to processing anterior and coronal dependently with known boundaries. Difficulty in assessing this feature arose mainly in the cases of NG, TH, and V. Bias probably accounts for the problem

with NG. Low spectral energy often caused TH to be estimated as a quiet region rather than a fricative; bias in the models for this broad class led to errors in assessing the feature anterior for TH. The problems for V are more difficult to interpret.

CORONAL The dependent modeling of anterior and coronal led to a d' which was 2% larger than that achieved through independent modeling of these features in the case of known boundaries. Difficulty for this feature arose mainly for the phonemes L, N, D, DH, and TH. The allophones EL and EN probably contribute to the problems with L and N. Low energy for DH and TH led these sounds to be classified as quiet and bias against +coronal led to errors for these sounds.

ROUND Difficulty for this feature arose mainly for the phonemes AA, UH, AH, and L. In general this feature is difficult with the local processing scheme of this chapter because the manifestation of this feature is a lowering of the formants relative to their position when rounding is not present rather than an absolute distinction between all +round and -round sounds.

TENSE Phonemes UW, OW, and AA led to the most difficulty for this feature. As with the case of round, the feature tense is manifest in more extreme formant positions than non-tense (lax) sounds but a universal distinction between tense and lax does not exist. longer in duration than lax sounds, however, more sophisticated durational models (rather than the exponential models implied by the Markov chain) would probably improve performance for this feature.

VOICE Difficulty in assessing the value of the feature voice arose primarily for the phonemes J and Z. The cessation of vocal fold vibration which is possible during the middle of the production of these sounds makes the voicing difficult to detect; incorporation of the information from the transition modeling of the next chapter is shown to improve d' significantly over that attained using local processing alone.

CONTINUANT Loss of boundary information caused a relatively large decrease in d' for the feature continuant, indicating the importance of transitional regions in assessing the value of this feature.

STRIDENT Errors in assessing the feature strident arose primarily for the phoneme V, most likely due to errors in the broad class estimation stage.

3.9 Weaknesses of the Local Processing Algorithm

The local processing algorithm has some shortcomings which be addressed through the complementary global processing of the next chapter.

One weakness of the local processing is that decisions about the feature composition of each frame are made based upon a small neighborhood around that frame. Information about spectra elsewhere in the waveform is incorporated into the attribute vector describing each frame only through the use of cepstral derivatives. Context effects are not modeled, and information outside of the TIMIT markings is not exploited in determining the feature composition within the region. Another weakness of the local processing scheme is that all frames within an estimated broad class portion representing a given feature state are assumed to have similar spectra. That is, for each estimated broad class portion, a unimodal Gaussian is used to model all frames corresponding to the presence of a given feature. This modeling technique implies a dichotomy of frames within each estimated broad class region so that the set of speech frames within each region fall into +feature and -feature half-planes. This assumption could be relaxed by replacing the unimodal Gaussian modeling with mixture models. Finally, the local processing does not fully model the interaction among features within each phoneme. Features are modeled within estimated broad class regions, thereby modeling the dependence of the secondary features discussed in Chapter 2 on the configuration of primary features. We have also investigated the joint modeling of place of articulation features within each broad class. However, the interaction between features may be more widespread.

4. A Distinctive Feature Representation – Global Processing

4.1 Introduction

The weaknesses of the the local processing discussed in section 3.9 are addressed through the complementary global processing of this chapter. To summarize the weaknesses of the local processing, we have that decisions about the feature composition of each frame are made based upon a small neighborhood around that frame. In addition, all frames within an estimated broad class portion representing a given feature state are assumed to have similar spectra, implying a separation of frames within each estimated broad class region into +feature and -feature half-planes. Finally, the local processing does not fully model the interaction among features.

The processing described in this chapter is complementary to that of the previous chapter in that it specifically addresses its weaknesses. This complementary processing takes into account the information present in the spectra associated with one sound in describing the feature composition of a neighboring sound by explicitly modeling transitional regions, resulting in a more global process than the procedure described in Chapter 3. Furthermore, in contrast to the bottom-up procedure of the previous chapter in which frames corresponding to the presence and to the absence of each feature were modeled directly within each estimated broad class portion, we use a top-down procedure in this chapter which relies on a mixture model rather

than a unimodal Gaussian model to assess feature probabilities. Models here are at the diphone level, and a mapping is performed to transform probabilities of phones into probabilities of features. This method captures the potential interdependence of features by modeling entire feature sets, or phonemes.

The method is shown to generally improve performance over using local processing alone in assessing the presence or absence of features.

4.2 Transition Modeling

The coarticulatory effects introduced in continuous speech warrant processing at a global level in order to incorporate information about a sound which appears in the spectra outside of the corresponding phone as well as the deformation of that phone due to feature spread across transitional regions.

The procedure described in this chapter relies on models of spectra in the neighborhood of transitions. To begin, we construct an attribute vector \bar{x} for a transition occurring at time t , where

$$\bar{x}(t) = \begin{bmatrix} \sum_{k=0}^2 x(t-k) \\ \sum_{k=1}^3 x(t+k) \end{bmatrix}$$

and x is a 28-dimensional vector consisting of 14 normalized cepstra and their time derivatives.

\mathcal{T} is defined to be the set of transitions in the data set; for training, as well as for testing in the case of a known segmentation, we take $\mathcal{T} = \{t : \phi(t) \neq \phi(t+1)\}$, where $\phi(t)$ is the TIMIT label occurring at frame t . For testing when the segmentation is not known, we take $\mathcal{T} = \{t : \hat{\sigma}(t) = 3, \hat{\sigma}(t+1) = 1\}$, where $\hat{\sigma}(t) \in \{1, 2, 3\}$ is the estimated portion of a broad class at time t .

4.3 Dimension Reduction

The limited number of training tokens of each transition mandates the use of reduced-order models. The problem of finding an appropriate reduced-order model given a finite number of training examples has two components. First, we must decide upon a system order which is small enough to allow for robust modeling in the face of limited training yet large enough to provide adequate representation of the attribute vectors. Having chosen a value for the order of the reduced system, we must then find a projection from the high-order to the low-order system. A class-specific transformation which captures principal components of all the data as well as those directions specific to an individual class, as outlined by Chernoff [3], will be considered.

For a given order p of our reduced system, we choose integers m and n such that $m + n = p$. Here m will represent the set of dimensions derived from the covariance data of the entire set of training tokens, and n will be the additional dimensions derived specifically for each transitional model.

To begin, we define the covariance matrix $\Sigma_{\alpha\beta}$ for transitions from frames representing primarily phoneme α to frames representing primarily phoneme β :

$$\begin{aligned} N_{\alpha\beta} &= \sum_{\{t:\phi(t)=\alpha,\phi(t+1)=\beta\}} 1 \\ \mu_{\alpha\beta} &= \frac{1}{N_{\alpha\beta}} \sum_{\{t:\phi(t)=\alpha,\phi(t+1)=\beta\}} \tilde{x}(t) \\ \Sigma_{\alpha\beta} &= \frac{1}{N_{\alpha\beta}} \sum_{\{t:\phi(t)=\alpha,\phi(t+1)=\beta\}} \tilde{x}(t)\tilde{x}(t)^T - \mu_{\alpha\beta}\mu_{\alpha\beta}^T \end{aligned}$$

and the “population covariance”

$$\Sigma = \sum_{\alpha,\beta} N_{\alpha\beta} \Sigma_{\alpha\beta}$$

Next, we define A_m to be the $P \times m$ matrix chosen so that the columns are the

eigenvectors associated with the m largest eigenvalues of the population covariance matrix Σ . We then perform the transformation $\Sigma_{\alpha\beta}^\perp = (I - \tilde{A}_m)^T \Sigma_{\alpha\beta}$ where \tilde{A}_m is a $P \times P$ matrix formed by adding $P - m$ columns of zeros to A_m . This transformation results in the projection of $\Sigma_{\alpha\beta}$ onto the nullspace of A_m .

Finally we define A_n to be the $P \times n$ matrix consisting of the n largest eigenvalues of $\Sigma_{\alpha\beta}^\perp$, and $A = [A_m | A_n]$.

In our experimentation we have observed that the value of m for a given p has little effect on performance in estimating the presence or absence of features. Therefore we have chosen $m = p$, $n = 0$, reducing the algorithm to one of principal component analysis. The results discussed in this chapter derive from a reduction of the original 56 dimensions to 15.

After performing the transformation $x' = A^T \tilde{x}$, we evaluate the Gaussian models for each possible phone transition. The following quantities are calculated for transitions from phoneme α to phoneme β for each $\alpha \neq \beta$:

$$\begin{aligned} \mu'_{\alpha\beta} &= \frac{1}{N_{\alpha\beta}} \sum_{\{t|\phi(t)=\alpha, \phi(t+1)=\beta\}} x'(t) = A^T \mu_{\alpha\beta} \\ \Sigma'_{\alpha\beta} &= \frac{1}{N_{\alpha\beta}} \sum_{\{t|\phi(t)=\alpha, \phi(t+1)=\beta\}} x'(t)x'(t)^T - \mu'_{\alpha\beta}\mu'_{\alpha\beta}{}^T = A^T \Sigma_{\alpha\beta} A \end{aligned}$$

4.4 Probability of Features from Transition Models

We shall say that $\alpha \ni f$ if the feature f is specified as “+” in the vector of features representing phoneme α . We have that the probability of feature f being encoded in the waveform in the left neighborhood of a transition at time t is:

$$\begin{aligned} p_t(f|x') &= \sum_{\alpha \ni f} p(\phi(t) = \alpha | x') \\ &= \sum_{\alpha \ni f} \sum_{\beta \neq \alpha} p(\phi(t) = \alpha, \phi(t+1) = \beta | x') \end{aligned}$$

$$= \frac{1}{p(x')} \sum_{\alpha \ni f} \sum_{\beta \neq \alpha} p(x' | \phi(t) = \alpha, \phi(t+1) = \beta) p(\phi(t) = \alpha, \phi(t+1) = \beta)$$

This procedure sums over all possible right contexts for a given phone, and then sums over all phones in which a given feature is present in order to arrive at the probability that a feature is present to the left of the transition. We can also assess this probability for the frames to the right of a transition by summing over left contexts. We have that the probability of feature f being encoded in the waveform in the right neighborhood of a transition at time r is:

$$\begin{aligned} p_r(f|x') &= \sum_{\alpha \ni f} p(\phi(r+1) = \alpha | x') \\ &= \sum_{\alpha \ni f} \sum_{\beta \neq \alpha} p(\phi(r) = \beta, \phi(r+1) = \alpha | x') \\ &= \frac{1}{p(x')} \sum_{\alpha \ni f} \sum_{\beta \neq \alpha} p(x' | \phi(r) = \beta, \phi(r+1) = \alpha) p(\phi(r) = \beta, \phi(r+1) = \alpha) \end{aligned}$$

The average of the probabilities derived from the left and right transitions provides a globally-derived estimate of the probability of feature f being encoded between the two transitions. That is, $\forall s : r \leq s < t, p_s(f|x') = \frac{1}{2}(p_{r+}(f|x') + p_{t-}(f|x'))$.

In order to combine this global estimate of the probability of a given feature being encoded in the waveform in the neighborhood of each frame with the local estimate described in the previous chapter, the average of the probabilities derived from these two processes is taken.

4.5 Results of Global Processing

Shown in Appendices 2 and 3 are the results of the global processing alone and of the combination of global with local processing, respectively. The cases of both known and unknown TIMIT boundary locations are considered. In this section, we summarize the results listed in the appendices and compare the performance with that

of the local processing of the previous chapter, relying upon the d' analysis technique described in section 3.8.1.

Table 4.1 indicates the resulting d' for each feature under various experimental conditions. The results of the local processing alone are reproduced here for easy comparison to the global processing. Conditions A-D of the table represent the case of known boundaries. Condition A represents the case of local processing in which all features are modeled independently. Condition B corresponds to local processing only, with dependent modeling of the features high, back, and low as well as anterior and coronal. In this experiment entries not shown are identical to the corresponding value in column A. Condition C corresponds to global processing alone, while case D represents the combined local and global processing.

Conditions E-G represent the case in which phonemic boundaries are unknown. Column E indicates the results using local processing alone, with dependent modeling of place features. Column F shows the results with global processing alone, and case G is the result of combining the outputs of scenarios E and F.

In table 4.2 is shown the change in performance between various scenarios of table 4.1. In the left column is the percent change in d' encountered in going from local only to combined local and global processing in the case of known boundaries. The center column of the table shows the same quantity in the case of unknown boundaries. In the right column of the table is shown the percent change in d' resulting from the combined local and global processing when boundary information is lost.

The agreement between local and global estimates of the probability of each feature when phonemic boundary information is given is investigated in table 4.3. The percent in the upper left of each matrix corresponds to the frequency with which both local and global probabilities are less than 0.5 thus rounding to zero in the VQ indexing scheme described in chapter 5. The number in the upper right corner of each matrix reflects the percentage of frames for which local processing led to a probability of feature present of less than 0.5 while global processing estimated the probability to be greater than 0.5. The lower left entry reflects the case in which the

local processing led to a probability greater than 0.5 but that resulting from global processing was less than 0.5. Finally, the lower right entry corresponds to the frames for which both the local and the global probability estimates were greater than 0.5. Shown in table 4.4 is the percentage of frames in which the local processing corresponds to a correctly-estimated feature value for the cases where either the locally- or the globally-derived probability is less than 0.5 while the other is greater than 0.5.

FEATURE	A	B	C	D	E	F	G
VOCALIC	3.807		3.628	3.921	3.010	2.906	3.047
STRIDENT	3.631		3.545	3.732	2.951	2.666	2.981
NASAL	3.503		3.137	3.499	2.910	2.558	2.937
CONSONANTAL	3.226		3.159	3.437	2.614	2.463	2.640
CONTINUANT	3.219		2.877	3.297	2.472	2.323	2.637
VOICE	2.758		2.928	3.050	2.440	2.563	2.711
ANTERIOR	2.705	2.666	2.495	2.832	2.119	1.914	2.191
ROUND	2.661		2.576	2.655	2.213	2.366	2.356
CORONAL	2.574	2.625	2.380	2.770	2.130	1.843	2.171
BACK	2.550	2.536	2.347	2.714	2.195	1.960	2.251
LOW	2.483	2.369	2.512	2.686	2.127	2.326	2.224
LABIAL	2.369		2.257	2.972	1.972	2.097	2.355
HIGH	2.278	2.442	2.111	2.451	2.038	1.745	2.136
TENSE	2.051		2.226	2.270	1.671	2.034	1.984

Table 4.1: Performance (d') for individual features in a variety of paradigms. **A:** Independent local, boundaries known. **B:** Dependent local, boundaries known. **C:** Global, boundaries known. **D:** Combined global and dependent local, boundaries known. **E:** Dependent local, boundaries unknown. **F:** Global, boundaries unknown. **G:** Combined global and dependent local, boundaries unknown.

Given below is a discussion of some anomalous results in the global processing as well as a comparison of the global to the local processing on an individual feature basis.

VOCALIC As was the case for local processing, the global processing had difficulty in assessing the value of the feature vocalic for the phonemes Y and W. The vowel-like structure of the spectra of these sounds as well as their relative infrequency led to the difficulties for these sounds. Combining the global information with the local did not improve the accuracy of determining the feature

FEATURE	$\Delta : B \rightarrow D$	$\Delta : E \rightarrow G$	$\Delta : D \rightarrow G$
VOCALIC	3.0%	1.2%	-22.3%
STRIDENT	2.8	1.0	-20.1
NASAL	-0.1	0.9	-16.1
CONSONANTAL	6.5	1.0	-23.2
CONTINUANT	2.4	6.7	-20.0
VOICE	10.6	11.1	-11.1
ANTERIOR	6.2	3.4	-22.6
ROUND	-0.2	6.5	-11.3
CORONAL	5.5	1.9	-21.6
BACK	7.0	2.6	-17.1
LOW	13.4	4.6	-17.2
LABIAL	25.5	19.4	-20.8
HIGH	0.4	4.8	-12.9
TENSE	10.7	18.7	-12.6

Table 4.2: **Left:** Percent change in overall d' for individual features in going from local to combined local and global processing when boundary information is given. **Center:** Percent change in d' in going from local to combined local and global processing when boundary information is not provided. **Right:** Percent change in d' for combined global and local processing when boundary information is lost.

vocalic for these two phonemes.

In terms of d' , combining local with global processing increased the overall separability of +vocalic and -vocalic sounds by 3% over local processing alone when boundary information was given and by 1.2% when not given, as seen in table 4.2. As shown in tables 4.3 and 4.4, the local and global estimates were consistent for 96% of the frames in the final test set with boundary information given. For the frames in which the two methods disagreed, the global estimate was correct 59% of the time. Referring again to table 4.2, we see that performance in assessing the presence of the feature vocalic degraded relatively severely when boundary information was lost, although d' remained larger for vocalic than for any other feature under this scenario.

CONSONANTAL Difficulty in assessing the value of the feature consonantal arose mainly in the cases of UW, UH, and W using the global technique. Confusion of these -consonantal sounds with the more frequent L, which is +consonantal,

VOCALIC		CONSONANTAL		HIGH		BACK	
0.59	0.02	0.28	0.05	0.73	0.10	0.81	0.03
0.02	0.37	0.02	0.65	0.04	0.13	0.07	0.09

LOW		ANTERIOR		CORONAL		NASAL	
0.87	0.04	0.63	0.10	0.64	0.08	0.92	0.02
0.04	0.05	0.05	0.22	0.06	0.22	0.02	0.05

LABIAL		ROUND		TENSE		VOICE	
0.73	0.05	0.81	0.02	0.73	0.05	0.37	0.04
0.15	0.07	0.11	0.06	0.12	0.10	0.07	0.52

CONTINUANT		STRIDENT	
0.11	0.03	0.82	0.01
0.04	0.82	0.03	0.14

Table 4.3: Agreement of local and global processing when boundaries are known. The upper left entry is the frequency with which both local and global probability estimates were less than 0.5. The upper right entry is the frequency with which the local probability was less than 0.5 but the global was greater than 0.5. The lower left entry is the frequency with which the local processing probability was greater than 0.5 but that of the global was less than 0.5. The lower right entry reflects the frequency with which both estimates were greater than 0.5.

FEATURE	%	FEATURE	%
VOCALIC	0.41	CONSONANTAL	0.44
HIGH	0.59	BACK	0.47
LOW	0.52	ANTERIOR	0.49
CORONAL	0.46	NASAL	0.49
LABIAL	0.27	ROUND	0.25
TENSE	0.31	VOICE	0.53
CONTINUANT	0.43	STRIDENT	0.39

Table 4.4: Percent of frames in which the local estimate is correct when local and global estimates disagree and boundaries are known.

was probably responsible for the difficulties.

The difficult cases of UH and L which arose from local processing are greatly improved through the use of combined local and global information.

Incorporation of global processing increased overall d' by 6.5% in the case of known boundaries and by 1.0% in the case of unknown boundaries, as indicated in table 4.2. Loss of boundary information was more devastating for the feature consonantal than for any other feature, causing a 23% decline in overall d' . These data highlight the importance of studying the transitional regions for assessing the feature consonantal as well as the sensitivity of the global processing for this feature to precise localization of the transitional region. As shown in tables 4.3 and 4.4, global and local processing disagreed for 7% of the frames in the final test set when boundary information was provided, with the global processing correct in 56% of those cases.

HIGH The global processing had difficulty with the feature high, in part because of the large bias toward -high across the entire set of phonemes. The relative infrequency, for example, of NG and ZH, along with their spectral similarity to N and Z, respectively, caused the frames associated with these phonemes to be wrongly judged to be -high. In the case of vowels, difficulty arose in the cases of UW, EY, IH, and UH. UW and UH were confused with the more frequent L while similarity of EY to the more frequent IY and of IH to EH led to the problems in these cases.

As indicated in table 4.2, including global information along with the local caused an increase in overall d' of 0.4% when boundary information was given and 4.8% when boundary locations were not provided. Loss of boundary information led to a decrease in overall d' by 12.9% for the combined local and global processing. As shown in tables 4.3 and 4.4, the dependent-local and global estimates were in agreement on 86% of the frames in the final test set in the case of known boundaries, with local processing correct in 59% of the disputed cases.

BACK As with the feature high, the global processing had a large bias toward -back

due to the relative infrequency of this feature across the entire set of phonemes. Particularly difficult to assess were the phonemes UW, UH, and NG. Again, the infrequency of NG led to problems for this phoneme while the similarity of UW and UH to L led to difficulties for these sounds.

Incorporation of global processing increased d' by 7.0% over local dependent processing alone when boundaries were known and 2.6% when they were unknown. Loss of boundary information caused a 17% decrease in overall d' for the combined local and global processing. As shown, local and global processing agreed for 90% of the frames in the final test set when boundary information was given, with global processing correct for 53% of the disagreements.

LOW Again, because of the low frequency of +low sounds, the global processing was strongly biased toward -low. Particularly difficult to assess were OW and AE presumably due to similarity with AO and AH, respectively.

With boundary information given, the incorporation of global processing led to an increase in d' of 13.4% over dependent local processing alone. When boundary locations were unknown, the combined global and local processing increased d' by 4.6% over the dependent local processing. The loss of boundary information led to a decline in d' of 17.2% for the combined local and global processing. Local and global estimates of the feature low agreed for 92% of the frames in the final test set when boundary information was known, with local processing correct in 52% of the disagreements.

ANTERIOR Difficulty in assessing the value of the feature anterior using global processing arose primarily in the cases of SH and ZH due to their relative infrequency and similarity to S and Z, respectively.

The combination of global processing with dependent local processing increased the overall d' for the feature anterior by 6.2% when boundary information was provided and by 3.4% when boundaries were unknown. Global and dependent-local estimates of the presence of the feature anterior agreed for 85% of the frames in the final test set when boundaries were known, with global processing

correct in 51% of the disputed cases. Loss of boundary information led to a decrease in overall d' of 22.6% for combined local and global processing, the second largest decrease of all features.

CORONAL The most difficult phonemes for which to correctly identify the feature coronal were TH and ZH. TH is often very low energy and therefore confused with silence, while the infrequency of ZH makes a model for this phoneme difficult to train.

The incorporation of global processing led to a 5.5% increase in d' over the use of dependent processing only in the case of known boundaries and 1.9% in the case of unknown boundaries. Global and dependent-local processing agreed for 86% of the frames in the final test set when boundary information was given, with global processing correct in 54% of the disagreements. Loss of boundary information led to a decrease in d' of 21.6% for the combined local and global processing.

ROUND The global processing is strongly biased toward -round due to the relative infrequency of this feature across the set of phonemes. Particularly difficult to assess, however, were UW and UH due to their confusability with L.

Incorporation of the global processing led to a decline in overall d' of 0.2% relative to local processing alone in the case of known boundaries when boundary information was provided but an increase of 6.5% when boundary locations were not given. In the case of known boundaries, local and global processing agreed for 87% of the frames in the final test set, and global processing was correct in 75% of the disputed cases, primarily due to the strong bias toward -low. Loss of boundary information caused a decrease in overall d' of 11.3% for the feature round, the second smallest decrease among the features.

TENSE Again the global processing is strongly biased toward -tense due to the infrequent occurrence of this feature across the full set of phonemes. Particular difficulty arose for the phonemes UW, OW and AA.

The use of global processing increased the overall d' for the feature tense by 10.7% over the local processing alone in the case of known boundaries and by 18.7% in the case of unknown boundaries. The local and global processing agreed for 83% of the frames in the final test set when boundary information was provided, with global processing correct in 69% of the disagreements. Loss of boundary information led to a decrease in overall d' of 12.6%.

VOICE The global processing had difficulty with the feature voice in the cases of G and J. A possible source of this error was the fact that information from the two edges of these phones was averaged in order to arrive at the global estimate, whereas the voicing evidence may be primarily evident at the release of the burst, especially in the case of J. Using the left edge alone could improve performance in assessing the feature voice for these phonemes.

The inclusion of global processing led to an increase in overall d' of 10.6% over the local processing alone for the feature voice when boundary locations were known and 11.1% when boundaries were not known. The loss of boundary information corresponded to a decrease in overall d' of 11.1% for the combined local and global processing, the smallest decrease of all features as indicated in table 4.2. Local and global estimates of the presence of voicing were in agreement for 89% of the frames in the final test set when boundaries were given; local processing was correct in 53% of the disputes.

CONTINUANT The most difficult phoneme for which to assess the value of the feature continuant was B. Presumably the very short nature of realizations of this phoneme as well as the use of information from both edges of these phones rather than just the left led to the problems here.

For the feature continuant, incorporation of global information increased the overall d' by 2.4% over that achieved by local processing alone when boundaries were known and by 6.7% when boundaries were unknown. Loss of boundary information led to a decrease in d' of 20% for the combined local and global processing. Local and global processing were in agreement for 93% of the frames

in the final test set when boundaries were known; global processing was correct in 57% of the disputes.

NASAL The use of global processing had little effect on the estimates of nasality; in the case of known boundaries d' declined by 0.1% over the use of local processing alone while in the case of unknown boundaries the combined processing led to an overall d' which was 0.9% higher than that achieved by local processing alone. Loss of boundary information led to a decrease of 16.1% in overall d' . With boundary information given, local and global estimates agreed for 97% of the frames in the final test set, with the global processing correct in 51% of the disputes.

STRIDENT Difficulty in assessing the value of the feature strident arose primarily for V, as was the case in the local processing. The fact that vocal tract resonances are often visible in the spectra for V causes confusability with the more common nasals N and M.

The incorporation of global processing led to an increase in overall d' of 2.8% over the use of local processing alone in the case of known boundaries. When boundary information was not provided, the use of global processing increased the overall d' by 1.0%. The loss of boundary information caused a decline in d' of 20.1% for the combined local and global processing. In the case of known boundaries, the local and global estimates agreed for 96% of the frames in the final test set, with the global processing correct in 61% of the disputes.

LABIAL The feature labial is redundant in the sense that sounds are +labial if and only if they are +anterior and -coronal. However, we found that phoneme identification was slightly better with the redundancy included.

The incorporation of the global processing led to an increase in overall d' of 25.5% over the use of local processing only in the case of known boundaries. When boundary information was not given, the use of global along with the local processing increased d' by 19.4% over that achieved through local processing

alone. Local and global processing differed in 20% of the frames in the final test in the case of known boundaries, with global processing correct in 73% of the disputed cases.

5. Phoneme Classification and Recognition

5.1 Phoneme Classification

5.1.1 Introduction

The feature representation of the waveform is developed with the ultimate goal of lexical access on the basis of features in the task of continuous speech recognition. However, rather than implementing such a system, we test the fidelity of our representation of the waveform on the intermediate tasks of phoneme classification and recognition.

The task of phoneme classification consists of, given the segmentation of the waveforms implied by the hand-marked TIMIT phoneme labels, identifying the phoneme label assigned to that segment. We use the procedure described in Chapter 3 in which boundary information is incorporated in order to derive the broad class estimates for each utterance in the test set. All of the results listed in this chapter rely on the combined local and global processing described in Chapters 3 and 4, where the local processing relies upon dependent modeling of the features high, back, and low as well as anterior and coronal.

Some work has been done by other researchers toward the use of linguistic feature representations for vowel classification. Meng [16] performed vowel classification experiments using neural network classifiers. That work investigated the use of an intermediate representation of vowels in terms of a set of distinctive linguistic features

versus the direct classification from the waveform. A slight decrease in performance by using the intermediate representation was observed. Leung & Zue [14] performed a similar experiment and found “quite similar” performance using a direct mapping to vowels or an intermediate feature representation.

These results illustrate that a speaker-independent feature representation is able to retain relevant information for performing classification. Our experiments have shown that an improvement in the accuracy of the feature representation leads to a direct improvement in phoneme classification accuracy.

5.1.2 Baseline Experiment: Gaussian Models

It is difficult to compare results of different phoneme classification experiments, as different training and testing sets, as well as different phoneme sets are used in virtually every experiment reported in the literature. Therefore, we conducted a baseline experiment with which to compare our results. Gaussian models of normalized cepstra and their derivatives were parameterized for each phoneme in our set using the TIMIT training set.

In order to arrive at an estimate for each phonemic region defined by the TIMIT labels, we evaluated the *a posteriori* probability of each phoneme given the observations x_1, \dots, x_T . Duration was modeled as third order Erlangian; the phoneme estimate was chosen as:

$$\begin{aligned} \hat{\phi} &= \operatorname{argmax}_{\phi} p(\phi | x_1, \dots, x_t, T) \\ &= \operatorname{argmax}_{\phi} p(T | \phi) p(\phi) \prod_{t=1}^T p(x_t | \phi) \\ &= \operatorname{argmax}_{\phi} \alpha_{\phi}^{T-3} (1 - \alpha_{\phi})^3 p(\phi) \prod_{t=1}^T p(x_t | \phi) \end{aligned}$$

The use Gaussian models of cepstral attributes trained on the development training set led to a phoneme classification rate of 58% on the development test set.

A different approach in which the observations are not assumed to be independent

but rather that $p(\phi|x_i) \approx p(\phi|x_j) \forall i, j \in \{1, \dots, T\}$ and that $p(\phi|x_1, \dots, x_T) \approx p(\phi|x_1)$ was also investigated. Under these assumptions the phoneme estimate is:

$$\begin{aligned} \hat{\phi} &= \operatorname{argmax}_{\phi} p(\phi|x_1, \dots, x_t, T) \\ &= \operatorname{argmax}_{\phi} p(T|\phi)p(\phi|x_1, \dots, x_T) \\ &= \operatorname{argmax}_{\phi} \alpha_{\phi}^{T-3} (1 - \alpha_{\phi})^3 \left(\prod_{t=1}^T p(\phi|x_t) \right)^{\frac{1}{T}} \end{aligned}$$

This revised approach, which still relies upon Gaussian models for the phoneme probabilities given the observations, resulted in a phoneme classification rate of 61% on our development test set.

5.1.3 Phoneme Classification from Feature Probabilities

In this section we describe a method of assigning a phoneme estimate to each phonemic segment of the waveform as defined by the TIMIT labels.

Assuming a segment has duration T , we choose our estimate of the identity of the underlying phoneme to be

$$\hat{\phi} = \operatorname{argmax}_{\phi} p(\phi|f_1, \dots, f_T, T)$$

where f_t is the vector of linguistic feature probabilities at time t .

To begin, we define $q_t : \mathfrak{R}^{14} \rightarrow Z^+ \cup \{0\}$ as:

$$q_t = \sum_{m \in \{1, \dots, 14 | f_m(t) > 0.5\}} 2^{m-1}$$

Thus, q_t serves as a quantizer for the real-valued feature probability vector f_t , so that $p(\phi|f) \approx p(\phi|q)$ where the right-hand side may be estimated as the number of occurrences of phoneme ϕ and index q relative to the total number of occurrences of

index q . Assuming that the duration and observed features are independent,

$$\begin{aligned}
\hat{\phi} &= \operatorname{argmax}_{\phi} p(\phi|f_1, \dots, f_T, T) \\
&= \operatorname{argmax}_{\phi} p(f_1, \dots, f_T, T|\phi)p(\phi) \\
&= \operatorname{argmax}_{\phi} p(f_1, \dots, f_T|\phi)p(\phi)p(T|\phi) \\
&= \operatorname{argmax}_{\phi} p(\phi|f_1, \dots, f_T)p(T|\phi) \\
&= \operatorname{argmax}_{\phi} p(\phi|q_1, \dots, q_T)p(T|\phi)
\end{aligned}$$

The first term on the right-hand side of the above is approximated by:

$$P(\phi|q_1, \dots, q_T) \approx \left(\prod_{t=1}^T p(\phi|q_t) \right)^{\frac{1}{T}}$$

Also, we model duration as a third order Erlang, so that:

$$p(T|\phi) = \alpha_{\phi}^{T-3}(1 - \alpha_{\phi})^3$$

where

$$\alpha_{\phi} = \frac{N_{\phi}}{3 + N_{\phi}}$$

and N_{ϕ} is the number of times a frame representing phoneme ϕ follows a frame representing the same phoneme in the training set.

5.1.4 Results of Phoneme Classification

For the phonemes listed in tables 2.2 through 2.4, a classification rate of 74.3% correct was achieved on the final test set. The correct phoneme was in the top two choices 85.6% of the time. Detailed confusion matrices are given in tables 5.1 through 5.3, with the phoneme presented listed vertically and the relative frequencies of responses listed horizontally.

	IY	UW	EY	OW	AA	IH	UH	EH	AH	AO	AE	Y
IY	0.91	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01
UW	0.10	0.14	0.00	0.10	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.00
EY	0.22	0.00	0.61	0.00	0.00	0.04	0.00	0.09	0.00	0.00	0.00	0.00
OW	0.00	0.00	0.00	0.54	0.04	0.04	0.00	0.00	0.00	0.21	0.00	0.00
AA	0.00	0.00	0.01	0.10	0.66	0.00	0.00	0.01	0.04	0.07	0.07	0.00
IH	0.29	0.01	0.05	0.02	0.00	0.47	0.00	0.11	0.01	0.00	0.02	0.00
UH	0.03	0.00	0.00	0.00	0.00	0.28	0.00	0.17	0.07	0.10	0.00	0.00
EH	0.01	0.00	0.02	0.01	0.00	0.11	0.00	0.54	0.03	0.05	0.16	0.00
AH	0.00	0.00	0.00	0.03	0.11	0.03	0.00	0.15	0.36	0.12	0.04	0.00
AO	0.00	0.00	0.00	0.08	0.12	0.00	0.00	0.00	0.05	0.67	0.00	0.00
AE	0.01	0.00	0.05	0.01	0.01	0.01	0.00	0.10	0.01	0.00	0.79	0.00
Y	0.36	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.44
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.06	0.00	0.00
L	0.02	0.00	0.01	0.00	0.01	0.02	0.00	0.02	0.01	0.05	0.00	0.00
R	0.01	0.00	0.01	0.00	0.00	0.05	0.00	0.03	0.01	0.01	0.00	0.00
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
NG	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
J	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Z	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ZH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.02
H#	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5.1: Relative frequency of confusions in phoneme classification. Phonemes presented are listed vertically and phonemes estimated are listed horizontally.

	W	L	R	N	M	NG	K	T	P	D	B	G
IY	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
UW	0.10	0.19	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EY	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OW	0.00	0.12	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AA	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IH	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UH	0.00	0.17	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
EH	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AH	0.00	0.01	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AO	0.04	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
W	0.70	0.10	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.00
L	0.03	0.70	0.03	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.01	0.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.00	0.02	0.01	0.76	0.09	0.00	0.00	0.00	0.00	0.00	0.01	0.00
M	0.01	0.01	0.01	0.15	0.68	0.00	0.00	0.00	0.00	0.00	0.03	0.00
NG	0.00	0.00	0.00	0.52	0.11	0.19	0.00	0.00	0.00	0.00	0.00	0.00
K	0.01	0.00	0.00	0.00	0.00	0.00	0.75	0.09	0.02	0.04	0.00	0.01
T	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.67	0.01	0.09	0.01	0.01
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.70	0.01	0.19	0.00
D	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.16	0.00	0.53	0.11	0.05
B	0.04	0.02	0.03	0.00	0.00	0.00	0.02	0.00	0.02	0.03	0.75	0.00
G	0.00	0.00	0.03	0.03	0.00	0.00	0.42	0.00	0.03	0.19	0.03	0.26
J	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.03	0.00	0.00
CH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00
F	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.01	0.00
S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.04	0.00	0.00	0.00
DH	0.00	0.01	0.01	0.12	0.02	0.00	0.00	0.12	0.00	0.11	0.03	0.02
Z	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ZH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V	0.02	0.03	0.05	0.03	0.05	0.00	0.00	0.00	0.00	0.00	0.02	0.00
H#	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5.2: Relative frequency of confusions in phoneme classification. Phonemes presented are listed vertically and phonemes estimated are listed horizontally.

	J	CH	F	S	SH	TH	DH	Z	ZH	V	H#
IY	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
UW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10
EY	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
OW	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
UH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
EH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
AO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06
L	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.01	0.08
R	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01
N	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.09
M	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.08
NG	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
K	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.06
T	0.01	0.05	0.01	0.01	0.00	0.00	0.02	0.00	0.00	0.00	0.05
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07
D	0.02	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.04
B	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.01	0.02
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
J	0.65	0.19	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00
CH	0.06	0.69	0.00	0.06	0.12	0.00	0.00	0.00	0.00	0.00	0.00
F	0.00	0.00	0.76	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.16
S	0.00	0.00	0.00	0.85	0.02	0.00	0.00	0.11	0.00	0.00	0.00
SH	0.00	0.04	0.04	0.12	0.78	0.00	0.00	0.00	0.00	0.00	0.02
TH	0.00	0.00	0.09	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.57
DH	0.00	0.00	0.00	0.01	0.00	0.00	0.24	0.03	0.00	0.00	0.25
Z	0.01	0.00	0.01	0.27	0.04	0.00	0.00	0.67	0.00	0.00	0.00
ZH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
V	0.00	0.00	0.06	0.00	0.00	0.00	0.03	0.06	0.00	0.43	0.19
H#	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.97

Table 5.3: Relative frequency of confusions in phoneme classification. Phonemes presented are listed vertically and phonemes estimated are listed horizontally.

5.2 Phoneme Recognition

The task of phoneme recognition consists of, given an utterance, listing the phoneme sequence which was produced. No boundary information is given.

This task has been addressed by various techniques, most notably hidden Markov models. Lee and Hon [12] reported accuracy on the recognition task under a variety of conditions. Their accuracies ranged from 58.8% for context independent recognition with no grammar to 73.8% for a bigram grammar and context-dependent modeling. Insertions were *not* counted as errors in their scoring algorithm.

5.2.1 Procedure

To perform this task we first estimate the linguistic features using the combined local and global processing, with the features anterior-coronal and high-back-low modeled dependently in the local processing.

Two means of deriving the estimated sequence of phonemes in a given utterance are described. The first uses a first-order Markov chain to model the trajectory through the phoneme state space for each utterance.

For an utterance of length N frames, our estimated phoneme string $\hat{\gamma}_1^N$ is:

$$\begin{aligned}\hat{\gamma}_1^N &= \arg \max_{\gamma_1^N} p(\gamma_1, \dots, \gamma_N | x_1, \dots, x_N) \\ &= \arg \max_{\gamma_1^N} \prod_{n=1}^N p(x_n | \gamma_n) p(\gamma_n | \gamma_{n-1}) \\ &= \arg \max_{\gamma_1^N} \prod_{n=1}^N \frac{p(\gamma_n | x_n)}{p(\gamma_n)} p(\gamma_n | \gamma_{n-1})\end{aligned}$$

where $p(\gamma_n | x_n)$ is the probability obtained from the vector quantization described in the previous section, $p(\gamma_n)$ is the prior probability of phoneme γ_n , $p(\gamma_n | \gamma_{n-1})$ is the transition probability of a frame representing γ_n following a frame chiefly representative of γ_{n-1} . We define $p(\gamma_n | \gamma_0) = \pi_n$, the probability of starting an utterance with phoneme n .

The second method of performing phoneme recognition assigns one phoneme label to each estimated broad class, thus forcing acceptance of the segmentation implied by the broad class processing. Dependence of the frames in the region is assumed. Again the VQ scheme described in the previous section is used to find the probability of each phoneme at each frame. In this scheme, however, the probabilities are smoothed by taking their average across the region. We combine this average with the three state exponential duration model for each candidate phoneme and choose as our estimate for the region the one having the highest score.

This procedure is identical to that used in phoneme classification, except that the implicit segmentation rather than the TIMIT markings is used to define phonemic regions. No grammar information, such as transition probabilities between phonemes, is used in this procedure. This method is shown to outperform the frame-based Markov chain method.

5.2.2 Scoring

We calculate \mathcal{A} , the accuracy [8] of the recognition, as:

$$\mathcal{A} = 1.0 - \frac{I + D + S}{N}$$

where I is the insertion rate, D is the deletion rate, S is the substitution rate, and N is the total number of symbols in the true label set.

A dynamic programming scheme is used to find the most favorable alignment of observation and truth for each sentence; symbols such as diphthongs which occurred in the truth but which we do not model are ignored. Thus, the occurrence of these symbols did not increase S or N .

After finding the best alignment using the above criterion, we calculate \mathcal{A} . Again we ignore substitutions involving excluded symbols, but here we also ignore insertions and deletions attributed to the excluded symbols.

5.2.3 Results of Phoneme Recognition

The Markov chain method led to an accuracy of 59% on our final test set. We achieved an accuracy of 64.6% by forcing the segmentation implied by the broad class stage. The errors for this latter experiment consisted of 2.6% insertions, 10.6% deletions, and 22.2% substitutions.

6. Discussion

6.1 Speech in the Context of General Pattern Recognition

As articulated by Mumford, Grenander has identified some general phenomena which arise in psychophysical pattern recognition problems [17]. Specifically, the problems of noise and blur, multi-scale phenomena, domain warping, and missing data surface in a variety of contexts. Particular instances of each of these general problems may be identified in the speech domain.

The problems of noise and blur are manifest in speech in variations among realizations of each phoneme due to speaker and channel as well as to coarticulation effects. We use statistical models to deal with noise, and mixture models of transitions to account for coarticulatory effects. As discussed in section 1.4.2, normalization of cepstra through spectral subtraction cancels static channel differences.

Variations in the durations among multiple pronunciations of a given sound constitute one instance of domain warping in the context of speech. This issue is addressed in our system through modeling feature trajectories with a Markov model.

The phenomenon of scale corresponds to the fact that the features underlying a given time frame are encoded in the waveform within a variety of time windows. Stevens' view of landmarks [25] fits into this framework, redefining the time axis in terms of events occurring on the segment rather than frame level. The problem of multi-scale phenomena has been addressed in this thesis through the joint use of local and global processing.

The fact that vocal tract structure is partially obscured by noise during the production of some consonants may be interpreted as an example of missing data; we have addressed this problem through explicitly modeling transitional regions.

6.2 The Use of Non-Linguistically-Motivated Features

In order to get a feel for the importance of the underlying linguistic motivation of feature sets, we have investigated the use of non-linguistically-motivated feature configurations for each phoneme. Specifically, we used the the feature assignments shown in tables 6.1- 6.3, which were obtained by randomly arranging the phoneme labels in tables 2.2- 2.4. Our models were retrained according to these feature assignments on the final training set. We then evaluated the probability of each feature being encoded near each frame of the final test set. These probabilities provided the input to the phoneme recognition system described in Chapter 5.

We achieved an accuracy of 21% (569 Insertions, 5884 Deletions, 12618 Substitutions out of 24116 tokens) on the phoneme recognition task using the randomized feature assignments. Thus, we conclude that the linguistic features provide a set of partitions of the acoustic space which are far superior to the randomly selected separations. This analysis, of course, does not demonstrate the superiority of linguistic features to statistically-selected separations through, for example, k-means techniques. However, it does suggest that the speech waveform exhibits physical correlates of the linguistically-motivated features which may be modeled reliably.

6.3 Potential Improvements Within The Existing Framework

We have developed a flexible framework, reflecting the inherent structure of speech, for the parameterization of the speech waveform in terms of linguistic features. Within

	VOWELS										
	R	F	P	G	AO	OW	JH	W	Z	CH	V
VOCALIC	+	+	+	+	+	+	+	+	+	+	+
CONSONANTAL	-	-	-	-	-	-	-	-	-	-	-
HIGH	+	+	-	-	-	+	+	-	-	-	-
BACK	-	+	-	+	+	-	+	-	+	+	-
LOW	-	-	-	+	+	-	-	-	-	-	+
ANTERIOR	-	-	-	-	-	-	-	-	-	-	-
CORONAL	-	-	-	-	-	-	-	-	-	-	-
ROUND	-	+	-	+	-	-	+	-	-	+	-
TENSE	+	+	+	+	+	-	-	-	-	-	-
VOICE	+	+	+	+	+	+	+	+	+	+	+
CONTINUANT	+	+	+	+	+	+	+	+	+	+	+
NASAL	-	-	-	-	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	-	-	-	-	-
LABIAL	-	-	-	-	-	-	-	-	-	-	-

Table 6.1: Assignment of non-linguistically-motivated features.

	GLIDES		LIQUIDS		NASALS			AFFRICATES		QUIET
	S	AH	TH	K	SH	EH	L	H#	B	D
VOCALIC	-	-	+	+	-	-	-	-	-	-
CONSONANTAL	-	-	+	+	+	+	+	+	+	+
HIGH	+	+	-	-	-	-	+	+	+	-
BACK	-	+	-	-	-	-	+	-	-	-
LOW	-	-	-	-	-	-	-	-	-	-
ANTERIOR	-	-	+	-	+	+	-	-	-	-
CORONAL	-	-	+	+	-	+	-	+	+	-
ROUND	-	+	-	-	-	-	-	-	-	-
TENSE	-	-	-	-	-	-	-	-	-	-
VOICE	+	+	+	+	+	+	+	-	+	-
CONTINUANT	+	+	+	+	-	-	-	-	-	+
NASAL	-	-	-	-	+	+	+	-	-	-
STRIDENT	-	-	-	-	-	-	-	+	+	-
LABIAL	-	-	-	-	+	-	-	-	-	-

Table 6.2: Assignment of non-linguistically-motivated features.

	PLOSIVES						FRICATIVES							
	M	UW	Y	NG	ZH	AE	T	EY	UH	IH	IY	N	DH	AA
VOCALIC	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CONSONANTAL	+	+	+	+	+	+	+	+	+	+	+	+	+	+
HIGH	-	-	+	-	-	+	-	-	-	-	-	-	+	+
BACK	-	-	+	-	-	+	-	-	-	-	-	-	-	-
LOW	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ANTERIOR	+	+	-	+	+	-	+	+	+	+	+	+	-	-
CORONAL	-	-	-	+	+	-	-	-	+	+	+	+	+	+
ROUND	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TENSE	-	-	-	-	-	-	-	-	-	-	-	-	-	-
VOICE	-	+	+	-	+	-	-	+	-	+	-	+	-	+
CONTINUANT	-	-	-	-	-	-	+	+	+	+	+	+	+	+
NASAL	-	-	-	-	-	-	-	-	-	-	-	-	-	-
STRIDENT	-	-	-	-	-	-	+	+	-	-	+	+	+	+
LABIAL	+	+	-	-	-	-	+	+	-	-	-	-	-	-

Table 6.3: Assignment of non-linguistically-motivated features.

that framework, a variety of modeling choices may be made. For example, we have built Gaussian models for each third of each broad class in order to accomplish a coarse grouping of the speech frames for subsequent feature analysis. The broad classes were derived from an analysis of the role of each feature in shaping the waveform. Once the underlying motivation for the broad class estimation stage is understood, however, one could use alternative clustering techniques such as k-means to achieve a segregation of the waveform space. This modeling choice could be easily adapted into the framework laid out in the thesis. Alternative techniques such as the use of mixture models or vector quantization for local processing could be incorporated as well. In the remainder of this section, we investigate several specific enhancements which could be made to the modeling choices described in Chapters 3 and 4.

6.3.1 Robust Global Processing

The transitional modeling described in Chapter 4 relies upon an accurate implicit segmentation by the broad class estimation stage. The global processing could be made more robust by the use of alternative or adapted modeling techniques.

An adaptation of the Gaussian model technique to address time-alignment might be to evaluate the Gaussians assuming that the change in estimated broad class fell at each frame within a small window of its actual position and sum the probabilities for each model within that window.

One example of an alternative, robust alignment procedure is that of time-delay neural networks [29]. Such structures could be easily adapted into our framework by using the implicit segment boundaries from the broad class processing to center a window, the frames in which are fed to the time-delay neural network. Each output node in the network would correspond to a phone-to-phone transition. Rather than choosing the output node with the largest total excitation over the window as an estimated phone-to-phone transition, however, we would take the relative excitations of each node to be the probability of the data given that transition. These probabilities would then replace the Gaussian scores and the algorithm would proceed as before.

6.3.2 Combining Local and Global Information

Combining the information arising from the global and local stages of processing was done by taking the average of the probability estimates derived from each method. In light of the analysis given in section 4.5, we note that certain features benefit differently from these two sources of information. Combining the estimates in a way which reflects the reliability of each probability estimate would presumably improve the accuracy of estimating features.

Furthermore, we have averaged the information from the left and the right edges of each region to assign a global probability to each frame in the interior of the region. Other methods of combining the information from the two edges which reflect the relative accuracy of the estimate derived from each edge could improve the global processing stage.

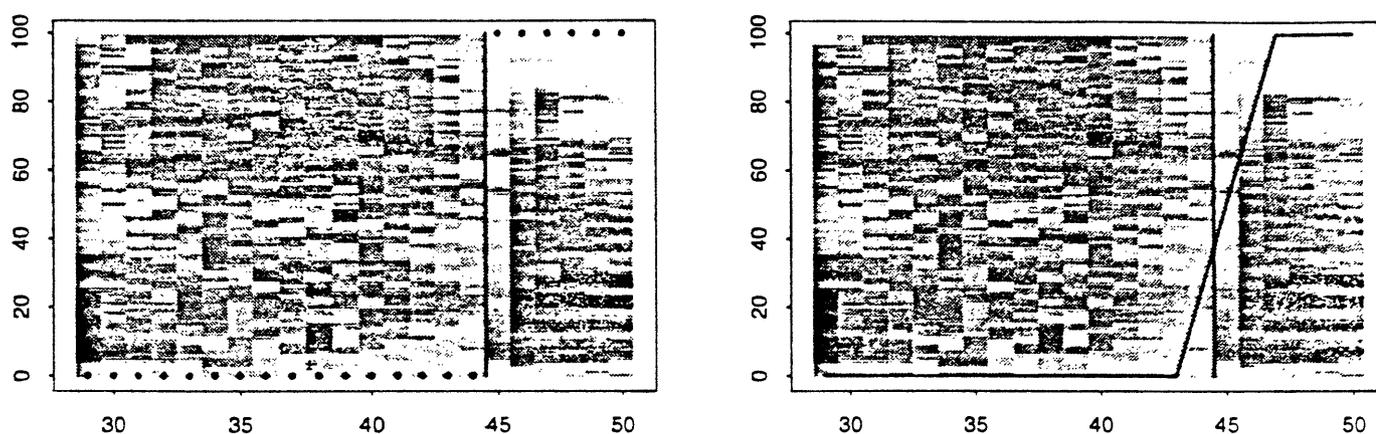


Figure 6.1: Spectrograms of the utterance “sh-ao” with truth labels for the feature “voice” superimposed.

6.3.3 Database Labels

We have used the TIMIT database for our experiments because of its availability, popularity, phonemic labels, and size. However, the manner in which the data are labeled may be a limitation on performance. Specifically, we have used a deterministic mapping from phoneme label to theoretical feature values to label the data for our experiments. This mapping corresponds to viewing an utterance as the output of a sequence of binary controls, where each control represents the presence or absence of a linguistic feature at each frame and all controls change state simultaneously at phonemic boundaries. This view is illustrated by the spectrogram in the left panel of figure 6.1. The utterance is “sh-ao” and the vertical line indicates the point at which the TIMIT label changes. The dots on the spectrogram illustrate the way we have labeled this token for the feature “voice.”

Physical manifestations of the features, however, do not follow the binary model, nor do changes in all features occur simultaneously at the “boundary,” as evidenced in part by the lack of voicing at the beginning of the vowel. One way to adjust our labeling scheme to reflect the fact that the TIMIT boundaries do not necessarily coincide with points in time at which a given feature changes state would be to use

real-valued truth labels, ramping between 0 and 1 across a transition as illustrated in the right panel of figure 6.1.

Another possibility for dealing with the labeling would be to adjust the truth labels based on running the training data through the system and using feature probabilities rather than theoretical values for subsequent training.

A database which is labeled according to a set of distinctive features actually produced is being amassed [24]. Our system is readily adaptable to such labels, which would result in clean training and presumably better performance in assessing feature probabilities.

6.4 A Potential Improvement Through Feedback

The idea of using feedback in speech processing is by no means a new one. Indeed, the theory of analysis-by-synthesis [28] involves a speech perception system in which hypothesized speech events are compared with an inventory of listener templates and phonological rules to adjust the perceived message. Implementation of such a scheme requires a catalog of phonological rules which govern distinctive feature trajectories.

Rather than implementing such rules, we consider modeling the simultaneous content of phoneme identity and speaker characteristics in the waveform through the use of feedback, thereby intertwining the tasks of speech recognition and speaker identification. The question of the level of abstraction at which the feedback is incorporated is a valid one; the following structure is but one layer in a potentially complex stratum. Referring to figure 6.2, feature analysis is performed first. The features are then used for estimating the sequence of phonemes uttered. The average attribute vectors associated with that sequence of phoneme estimates are subtracted from the attribute vectors parameterizing the waveform. The resulting signal is rich in speaker characteristics; the residual vectors are averaged and the original waveform is then normalized by these “characteristics of the speaker” and the features are reestimated. This structure would enable the use of feedback at low levels of abstraction without requiring knowledge of vocabularies and syntax.

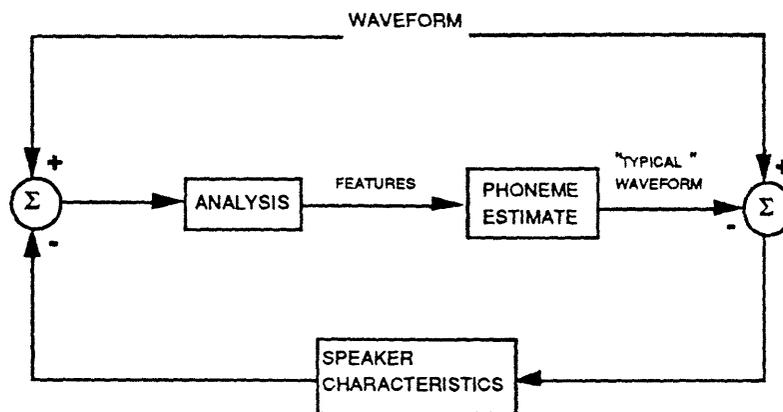


Figure 6.2: Block diagram for the incorporation of feedback to adjust estimates based on speaker independent properties of the waveform.

6.5 Summary

In this thesis we have described a method of parameterizing the speech waveform in terms of a set of linguistic features. The structure of the procedure reflects the hierarchical encoding of linguistic features in the speech waveform. Furthermore, the processing proceeds at different time scales with the global processing complementing the local by addressing some of its specific weaknesses.

The theoretical framework, constructed on the basis of a careful consideration of the physical phenomena encompassed in the speech signal, is flexible enough to allow for potential improvements through the use of more sophisticated modeling choices.

We have demonstrated the ability to effectively perform phoneme classification and recognition on the basis of the feature parameterization.

Finally, representation of the waveform in terms of linguistic features allows for lexical access at the feature level, as discussed in the opening chapter.

7. Appendices

7.1 Individual Results – Local Processing

Shown in the tables of this section are results of linguistic feature estimation for individual phonemes, using the local processing scheme described in Chapter 3. Entries reflect the relative frequency of frames representing a given phoneme estimated to represent the presence of a given feature. The “+” or “-” after each entry signify the theoretical presence or absence of the feature for that phoneme. Thus, ideal performance would be for each entry followed by a “+” to be 1.000 and each entry followed by a “-” to be 0.000. Tables 7.1 through 7.3 reflect performance in the case of known TIMIT boundaries. We show the results for features vocalic through coronal in one table and round through labial in a second for readability only. Tables 7.1 and 7.2 compare the dependent and independent processing of the features high, back, and low as well as anterior and coronal. Tables 7.4 and 7.5 indicate the performance in the case of unknown TIMIT markings.

	VOCALIC	CONSON.	HIGH	BACK	LOW	ANTERIOR	CORONAL
IY	0.934+	0.050-	0.874+	0.003-	0.000-	0.015-	0.015-
UW	0.973+	0.266-	0.471+	0.563+	0.000-	0.151-	0.200-
EY	0.997+	0.003-	0.207-	0.000-	0.000-	0.003-	0.001-
OW	0.998+	0.106-	0.000-	0.848+	0.323+	0.153-	0.194-
AA	0.999+	0.012-	0.000-	0.916+	0.678+	0.001-	0.005-
IH	0.984+	0.045-	0.575+	0.074-	0.022-	0.010-	0.032-
UH	0.993+	0.418-	0.293+	0.254+	0.000-	0.160-	0.269-
EH	0.997+	0.066-	0.042-	0.132-	0.190-	0.002-	0.028-
AH	0.962+	0.083-	0.012-	0.546+	0.075-	0.016-	0.023-
AO	0.981+	0.060-	0.008-	0.952+	0.150-	0.025-	0.052-
AE	0.998+	0.002-	0.001-	0.031-	0.700+	0.001-	0.000-
Y	0.404-	0.172-	0.775+	0.000-	0.000-	0.004-	0.060-
W	0.236-	0.155-	0.680+	0.859+	0.001-	0.034-	0.026-
L	0.847+	0.643+	0.039-	0.287-	0.056-	0.648+	0.615+
R	0.965+	0.828+	0.064-	0.051-	0.008-	0.010-	0.740+
N	0.052-	0.958+	0.007-	0.012-	0.005-	0.770+	0.571+
M	0.048-	0.981+	0.004-	0.011-	0.004-	0.870+	0.070-
NG	0.162-	0.822+	0.384+	0.254+	0.000-	0.597-	0.289-
K	0.000-	1.000+	0.565+	0.560+	0.000-	0.092-	0.066-
T	0.000-	1.000+	0.101-	0.021-	0.000-	0.677+	0.684+
P	0.000-	1.000+	0.009-	0.009-	0.000-	0.674+	0.005-
D	0.008-	0.984+	0.106-	0.023-	0.000-	0.624+	0.588+
B	0.059-	1.000+	0.021-	0.021-	0.000-	0.582+	0.038-
G	0.000-	1.000+	0.620+	0.598+	0.000-	0.145-	0.056-
J	0.000-	1.000+	0.927+	0.000-	0.000-	0.045-	0.952+
CH	0.000-	1.000+	0.908+	0.000-	0.000-	0.085-	0.993+
F	0.000-	1.000+	0.029-	0.015-	0.000-	0.818+	0.038-
S	0.002-	0.999+	0.033-	0.000-	0.000-	0.940+	0.985+
SH	0.012-	0.988+	0.887+	0.000-	0.000-	0.092-	0.969+
TH	0.002-	0.998+	0.000-	0.002-	0.002-	0.600+	0.252+
DH	0.029-	0.996+	0.018-	0.018-	0.000-	0.701+	0.434+
Z	0.002-	0.998+	0.068-	0.001-	0.000-	0.899+	0.982+
ZH	0.000-	1.000+	0.000+	0.000-	0.000-	0.000-	1.000+
V	0.134-	0.895+	0.064-	0.057-	0.001-	0.582+	0.119-
H#	0.004-	0.996+	0.001-	0.001-	0.000-	0.018-	0.007-

Table 7.1: Performance for individual phonemes using local processing with known boundaries when the features high-back-low and anterior-coronal are modeled dependently.

	VOCALIC	CONSON.	HIGH	BACK	LOW	ANTERIOR	CORONAL
IY	0.934+	0.050-	0.920+	0.019-	0.000-	0.018-	0.028-
UW	0.973+	0.266-	0.315+	0.597+	0.056-	0.126-	0.167-
EY	0.997+	0.003-	0.409-	0.000-	0.026-	0.003-	0.001-
OW	0.998+	0.106-	0.000-	0.970+	0.883+	0.229-	0.098-
AA	0.999+	0.012-	0.001-	0.968+	0.972+	0.001-	0.005-
IH	0.984+	0.045-	0.595+	0.119-	0.050-	0.019-	0.044-
UH	0.993+	0.418-	0.204+	0.551+	0.177-	0.160-	0.232-
EH	0.997+	0.066-	0.077-	0.197-	0.335-	0.002-	0.055-
AH	0.962+	0.083-	0.001-	0.777+	0.573-	0.028-	0.034-
AO	0.981+	0.060-	0.001-	0.969+	0.695-	0.033-	0.058-
AE	0.998+	0.002-	0.001-	0.168-	0.699+	0.001-	0.000-
Y	0.404-	0.172-	0.805+	0.000-	0.000-	0.004-	0.090-
W	0.236-	0.155-	0.577+	0.882+	0.068-	0.055-	0.041-
L	0.847+	0.643+	0.070-	0.612-	0.389-	0.714+	0.442+
R	0.965+	0.828+	0.073-	0.117-	0.061-	0.006-	0.811+
N	0.052-	0.958+	0.015-	0.117-	0.009-	0.862+	0.675+
M	0.048-	0.981+	0.009-	0.097-	0.005-	0.906+	0.142-
NG	0.162-	0.822+	0.260+	0.483+	0.003-	0.597-	0.330-
K	0.000-	1.000+	0.755+	0.865+	0.000-	0.124-	0.091-
T	0.000-	1.000+	0.161-	0.058-	0.000-	0.729+	0.771+
P	0.000-	1.000+	0.036-	0.095-	0.000-	0.750+	0.065-
D	0.008-	0.984+	0.166-	0.114-	0.000-	0.715+	0.661+
B	0.059-	1.000+	0.003-	0.082-	0.000-	0.774+	0.135-
G	0.000-	1.000+	0.676+	0.743+	0.000-	0.190-	0.106-
J	0.000-	1.000+	0.949+	0.000-	0.000-	0.045-	0.969+
CH	0.000-	1.000+	0.908+	0.000-	0.000-	0.085-	0.993+
F	0.000-	1.000+	0.049-	0.023-	0.000-	0.793+	0.090-
S	0.002-	0.999+	0.061-	0.000-	0.000-	0.958+	0.993+
SH	0.012-	0.988+	0.887+	0.000-	0.000-	0.148-	0.884+
TH	0.002-	0.998+	0.000-	0.002-	0.002-	0.600+	0.286+
DH	0.029-	0.996+	0.024-	0.032-	0.000-	0.776+	0.562+
Z	0.002-	0.998+	0.073-	0.002-	0.000-	0.911+	0.997+
ZH	0.000-	1.000+	0.000+	0.000-	0.000-	1.000-	1.000+
V	0.134-	0.895+	0.052-	0.086-	0.010-	0.550+	0.098-
H#	0.004-	0.996+	0.003-	0.011-	0.001-	0.019-	0.010-

Table 7.2: Performance using local processing when phonemic boundaries are known and all features are modeled independently.

	ROUND	TENSE	VOICE	CONTIN.	NASAL	STRIDENT	LABIAL
IY	0.006-	0.943+	0.982+	0.968+	0.021-	0.006-	0.005-
UW	0.755+	0.570+	0.998+	0.995+	0.002-	0.000-	0.113-
EY	0.000-	0.812+	0.999+	0.999+	0.001-	0.001-	0.001-
OW	0.970+	0.551+	0.999+	0.998+	0.001-	0.000-	0.001-
AA	0.456-	0.315+	0.999+	0.999+	0.000-	0.001-	0.023-
IH	0.114-	0.326-	0.998+	0.992+	0.007-	0.001-	0.006-
UH	0.488+	0.256-	1.000+	0.996+	0.000-	0.000-	0.004-
EH	0.134-	0.070-	0.998+	0.998+	0.001-	0.001-	0.001-
AH	0.529-	0.099-	0.965+	0.997+	0.002-	0.000-	0.010-
AO	0.910+	0.237-	0.998+	0.999+	0.000-	0.001-	0.008-
AE	0.049-	0.084-	0.999+	0.999+	0.000-	0.000-	0.001-
Y	0.000-	0.633-	0.985+	0.996+	0.004-	0.000-	0.022-
W	0.863+	0.040-	0.990+	0.987+	0.000-	0.000-	0.054-
L	0.582-	0.350-	0.998+	0.972+	0.026-	0.000-	0.047-
R	0.082-	0.089-	0.986+	0.993+	0.002-	0.001-	0.066-
N	0.025-	0.016-	0.996+	0.153-	0.844+	0.004-	0.170-
M	0.032-	0.014-	0.983+	0.115-	0.874+	0.005-	0.815+
NG	0.000-	0.197-	1.000+	0.213-	0.819+	0.000-	0.140-
K	0.000-	0.000-	0.093-	0.109-	0.000-	0.046-	0.033-
T	0.000-	0.000-	0.124-	0.098-	0.000-	0.138-	0.040-
P	0.000-	0.000-	0.140-	0.112-	0.000-	0.014-	0.811+
D	0.000-	0.000-	0.650+	0.215-	0.000-	0.161-	0.119-
B	0.118-	0.006-	0.912+	0.291-	0.003-	0.000-	0.782+
G	0.000-	0.000-	0.654+	0.039-	0.000-	0.000-	0.017-
J	0.000-	0.000-	0.579+	0.073-	0.000-	0.915+	0.000-
CH	0.000-	0.000-	0.000-	0.195-	0.000-	0.966+	0.000-
F	0.000-	0.000-	0.023-	0.958+	0.000-	0.774+	0.941+
S	0.000-	0.000-	0.155-	0.986+	0.000-	0.997+	0.025-
SH	0.000-	0.001-	0.012-	0.936+	0.000-	0.988+	0.102-
TH	0.002-	0.000-	0.255-	0.900+	0.000-	0.191-	0.314-
DH	0.018-	0.037-	0.709+	0.638+	0.157-	0.066-	0.265-
Z	0.001-	0.001-	0.630+	0.994+	0.002-	0.986+	0.037-
ZH	0.000-	0.000-	1.000+	1.000+	0.000-	0.000+	0.000-
V	0.103-	0.050-	0.833+	0.940+	0.054-	0.371+	0.784+
H#	0.001-	0.006-	0.151-	0.990+	0.008-	0.004-	0.127-

Table 7.3: Performance using local processing when phonemic boundaries are known.

	VOCALIC	CONSON.	HIGH	BACK	LOW	ANTERIOR	CORONAL
IY	0.870+	0.124-	0.823+	0.025-	0.006-	0.044-	0.059-
UW	0.849+	0.293-	0.532+	0.595+	0.050-	0.191-	0.216-
EY	0.952+	0.064-	0.295-	0.014-	0.022-	0.034-	0.032-
OW	0.931+	0.210-	0.032-	0.763+	0.450+	0.252-	0.264-
AA	0.964+	0.068-	0.008-	0.822+	0.633+	0.032-	0.049-
IH	0.885+	0.124-	0.554+	0.069-	0.059-	0.063-	0.094-
UH	0.869+	0.385-	0.311+	0.315+	0.004-	0.184-	0.254-
EH	0.921+	0.134-	0.077-	0.128-	0.221-	0.066-	0.104-
AH	0.902+	0.149-	0.038-	0.587+	0.161-	0.079-	0.106-
AO	0.932+	0.117-	0.049-	0.852+	0.226-	0.078-	0.123-
AE	0.959+	0.044-	0.028-	0.047-	0.671+	0.019-	0.018-
Y	0.476-	0.161-	0.835+	0.000-	0.000-	0.034-	0.101-
W	0.277-	0.165-	0.630+	0.824+	0.036-	0.088-	0.064-
L	0.833+	0.549+	0.081-	0.380-	0.130-	0.573+	0.568+
R	0.915+	0.772+	0.082-	0.086-	0.025-	0.044-	0.706+
N	0.081-	0.920+	0.043-	0.029-	0.010-	0.784+	0.616+
M	0.077-	0.942+	0.022-	0.037-	0.008-	0.838+	0.123-
NG	0.210-	0.787+	0.343+	0.225+	0.000-	0.584-	0.349-
K	0.013-	0.985+	0.558+	0.554+	0.005-	0.124-	0.095-
T	0.016-	0.986+	0.154-	0.029-	0.002-	0.629+	0.734+
P	0.048-	0.966+	0.064-	0.047-	0.009-	0.614+	0.062-
D	0.057-	0.951+	0.171-	0.078-	0.000-	0.570+	0.611+
B	0.162-	0.876+	0.071-	0.079-	0.012-	0.579+	0.115-
G	0.067-	0.972+	0.637+	0.575+	0.000-	0.151-	0.084-
J	0.006-	1.000+	0.898+	0.006-	0.000-	0.042-	0.944+
CH	0.000-	1.000+	0.833+	0.000-	0.000-	0.130-	0.973+
F	0.032-	0.972+	0.042-	0.033-	0.005-	0.801+	0.080-
S	0.003-	0.998+	0.054-	0.002-	0.001-	0.908+	0.947+
SH	0.020-	0.983+	0.798+	0.000-	0.000-	0.145-	0.930+
TH	0.034-	0.970+	0.027-	0.011-	0.007-	0.582+	0.330+
DH	0.121-	0.903+	0.066-	0.031-	0.001-	0.622+	0.444+
Z	0.011-	0.991+	0.062-	0.003-	0.000-	0.871+	0.913+
ZH	0.286-	0.714+	0.000+	0.000-	0.000-	0.286-	0.429+
V	0.325-	0.774+	0.120-	0.114-	0.018-	0.492+	0.177-
H#	0.019-	0.985+	0.020-	0.010-	0.001-	0.069-	0.043-

Table 7.4: Performance using local processing when phonemic boundaries are unknown. High-back-low and anterior-coronal are modeled dependently.

	ROUND	TENSE	VOICE	CONTIN.	NASAL	STRIDENT	LABIAL
IY	0.024-	0.862+	0.958+	0.952+	0.028-	0.025-	0.051-
UW	0.633+	0.615+	0.980+	0.962+	0.027-	0.007-	0.203-
EY	0.013-	0.678+	0.987+	0.981+	0.015-	0.013-	0.077-
OW	0.844+	0.419+	0.990+	0.964+	0.031-	0.015-	0.116-
AA	0.473-	0.362+	0.982+	0.985+	0.005-	0.006-	0.074-
IH	0.095-	0.292-	0.983+	0.962+	0.017-	0.048-	0.082-
UH	0.532+	0.230-	0.987+	0.967+	0.011-	0.020-	0.090-
EH	0.121-	0.092-	0.981+	0.962+	0.029-	0.019-	0.068-
AH	0.453-	0.125-	0.968+	0.971+	0.023-	0.036-	0.103-
AO	0.856+	0.291-	0.978+	0.988+	0.006-	0.012-	0.057-
AE	0.035-	0.106-	0.979+	0.981+	0.012-	0.005-	0.035-
Y	0.034-	0.652-	1.000+	0.921+	0.037-	0.026-	0.037-
W	0.837+	0.126-	0.978+	0.967+	0.025-	0.008-	0.185-
L	0.597-	0.389-	0.976+	0.951+	0.032-	0.010-	0.093-
R	0.108-	0.110-	0.964+	0.976+	0.010-	0.020-	0.123-
N	0.069-	0.074-	0.973+	0.205-	0.775+	0.032-	0.245-
M	0.135-	0.086-	0.969+	0.172-	0.819+	0.013-	0.792+
NG	0.149-	0.222-	0.990+	0.273-	0.727+	0.000-	0.203-
K	0.040-	0.054-	0.079-	0.139-	0.000-	0.045-	0.119-
T	0.010-	0.029-	0.139-	0.188-	0.002-	0.195-	0.079-
P	0.071-	0.056-	0.214-	0.253-	0.000-	0.042-	0.687+
D	0.021-	0.047-	0.598+	0.290-	0.028-	0.148-	0.158-
B	0.162-	0.109-	0.871+	0.394-	0.029-	0.015-	0.694+
G	0.050-	0.017-	0.553+	0.168-	0.011-	0.022-	0.084-
J	0.006-	0.008-	0.497+	0.201-	0.000-	0.924+	0.020-
CH	0.027-	0.000-	0.058-	0.290-	0.000-	0.935+	0.085-
F	0.053-	0.029-	0.095-	0.901+	0.001-	0.688+	0.863+
S	0.019-	0.011-	0.166-	0.976+	0.000-	0.976+	0.038-
SH	0.010-	0.009-	0.051-	0.848+	0.000-	0.965+	0.137-
TH	0.011-	0.009-	0.220-	0.816+	0.005-	0.277-	0.384-
DH	0.065-	0.063-	0.697+	0.647+	0.154-	0.093-	0.222-
Z	0.012-	0.020-	0.554+	0.956+	0.006-	0.947+	0.052-
ZH	0.000-	0.000-	0.714+	0.857+	0.000-	0.286+	0.000-
V	0.201-	0.139-	0.812+	0.904+	0.071-	0.268+	0.695+
H#	0.020-	0.031-	0.184-	0.945+	0.019-	0.036-	0.154-

Table 7.5: Performance using local processing when phonemic boundaries are unknown.

7.2 Individual Results – Global Processing

Shown in the tables of this section are results of linguistic feature estimation for individual phonemes, using the global processing scheme described in Chapter 4. Tables 7.6 and 7.7 reflect performance in the case of known TIMIT boundaries, while tables 7.8 and 7.9 indicate the performance in the case of unknown TIMIT markings.

	VOCALIC	CONSON.	HIGH	BACK	LOW	ANTERIOR	CORONAL
IY	0.953+	0.040-	0.864+	0.000-	0.000-	0.014-	0.007-
UW	0.995+	0.374-	0.315+	0.243+	0.000-	0.162-	0.264-
EY	0.980+	0.036-	0.244-	0.000-	0.000-	0.020-	0.020-
OW	0.998+	0.235-	0.017-	0.491+	0.133+	0.129-	0.157-
AA	0.999+	0.032-	0.001-	0.704+	0.365+	0.001-	0.021-
IH	0.966+	0.063-	0.495+	0.008-	0.000-	0.017-	0.021-
UH	0.919+	0.313-	0.120+	0.230+	0.000-	0.088-	0.125-
EH	0.963+	0.111-	0.126-	0.047-	0.081-	0.016-	0.052-
AH	0.997+	0.127-	0.001-	0.478+	0.102-	0.002-	0.125-
AO	0.957+	0.133-	0.007-	0.704+	0.049-	0.058-	0.080-
AE	0.970+	0.053-	0.065-	0.012-	0.157+	0.002-	0.002-
Y	0.727-	0.180-	0.745+	0.000-	0.000-	0.004-	0.060-
W	0.342-	0.362-	0.486+	0.598+	0.000-	0.128-	0.202-
L	0.803+	0.847+	0.023-	0.120-	0.000-	0.698+	0.694+
R	0.958+	0.867+	0.027-	0.018-	0.000-	0.012-	0.780+
N	0.063-	0.965+	0.003-	0.002-	0.000-	0.822+	0.606+
M	0.039-	0.976+	0.007-	0.014-	0.000-	0.851+	0.220-
NG	0.057-	0.943+	0.098+	0.003+	0.000-	0.708-	0.251-
K	0.000-	1.000+	0.632+	0.614+	0.000-	0.231-	0.137-
T	0.001-	0.999+	0.122-	0.044-	0.000-	0.773+	0.606+
P	0.003-	0.997+	0.040-	0.028-	0.000-	0.837+	0.045-
D	0.041-	0.959+	0.060-	0.026-	0.000-	0.723+	0.627+
B	0.115-	0.968+	0.024-	0.050-	0.006-	0.659+	0.091-
G	0.017-	0.983+	0.486+	0.514+	0.000-	0.179-	0.134-
J	0.000-	1.000+	0.653+	0.000-	0.000-	0.144-	0.969+
CH	0.000-	1.000+	0.546+	0.000-	0.000-	0.184-	0.993+
F	0.000-	1.000+	0.013-	0.000-	0.000-	0.848+	0.116-
S	0.001-	0.999+	0.015-	0.000-	0.000-	0.983+	0.958+
SH	0.002-	0.998+	0.477+	0.000-	0.000-	0.426-	0.924+
TH	0.002-	1.000+	0.000-	0.000-	0.000-	0.602+	0.436+
DH	0.046-	0.969+	0.022-	0.018-	0.000-	0.729+	0.509+
Z	0.002-	0.999+	0.043-	0.000-	0.000-	0.937+	0.999+
ZH	0.000-	1.000+	0.000+	0.000-	0.000-	1.000-	0.000+
V	0.057-	0.997+	0.001-	0.000-	0.000-	0.759+	0.163-
H#	0.002-	0.998+	0.003-	0.003-	0.000-	0.059-	0.022-

Table 7.6: Performance of estimating features for individual phonemes when phonemic boundaries are known, using global processing at the left and right boundaries of each phone. The global estimate is derived from the arithmetic average of the probabilities at the two edges.

	ROUND	TENSE	VOICE	CONTIN.	NASAL	STRIDENT	LABIAL
IY	0.000-	0.772+	0.994+	0.992+	0.007-	0.001-	0.000-
UW	0.074+	0.083+	1.000+	0.995+	0.002-	0.000-	0.000-
EY	0.000-	0.501+	0.999+	0.999+	0.001-	0.001-	0.000-
OW	0.228+	0.133+	1.000+	0.999+	0.001-	0.000-	0.000-
AA	0.031-	0.243+	0.999+	0.999+	0.000-	0.001-	0.000-
IH	0.000-	0.089-	0.999+	0.994+	0.001-	0.001-	0.000-
UH	0.182+	0.094-	1.000+	0.998+	0.000-	0.000-	0.000-
EH	0.004-	0.036-	0.999+	0.984+	0.008-	0.000-	0.001-
AH	0.014-	0.074-	1.000+	0.999+	0.000-	0.000-	0.000-
AO	0.399+	0.036-	0.999+	0.979+	0.000-	0.001-	0.000-
AE	0.000-	0.056-	1.000+	0.999+	0.000-	0.000-	0.000-
Y	0.000-	0.292-	0.985+	0.996+	0.000-	0.000-	0.000-
W	0.553+	0.000-	0.992+	0.995+	0.000-	0.000-	0.004-
L	0.076-	0.000-	0.999+	0.970+	0.018-	0.000-	0.005-
R	0.004-	0.003-	0.991+	0.994+	0.005-	0.000-	0.002-
N	0.000-	0.013-	0.948+	0.236-	0.726+	0.000-	0.039-
M	0.007-	0.000-	0.972+	0.239-	0.735+	0.000-	0.385+
NG	0.000-	0.044-	0.927+	0.225-	0.775+	0.000-	0.022-
K	0.000-	0.000-	0.117-	0.044-	0.000-	0.017-	0.021-
T	0.000-	0.000-	0.135-	0.150-	0.000-	0.109-	0.009-
P	0.000-	0.000-	0.287-	0.140-	0.000-	0.009-	0.622+
D	0.000-	0.005-	0.674+	0.158-	0.000-	0.101-	0.057-
B	0.003-	0.000-	0.806+	0.432-	0.000-	0.000-	0.312+
G	0.000-	0.000-	0.609+	0.134-	0.017-	0.011-	0.045-
J	0.000-	0.000-	0.520+	0.181-	0.000-	0.794+	0.000-
CH	0.000-	0.000-	0.246-	0.232-	0.000-	0.901+	0.000-
F	0.000-	0.000-	0.051-	0.989+	0.000-	0.646+	0.553+
S	0.000-	0.000-	0.259-	0.988+	0.000-	0.971+	0.007-
SH	0.000-	0.000-	0.087-	0.898+	0.000-	0.945+	0.043-
TH	0.000-	0.000-	0.100-	0.939+	0.000-	0.243-	0.100-
DH	0.018-	0.001-	0.672+	0.790+	0.031-	0.018-	0.012-
Z	0.000-	0.000-	0.775+	1.000+	0.000-	0.975+	0.000-
ZH	0.000-	0.000-	1.000+	1.000+	0.000-	1.000+	0.000-
V	0.000-	0.000-	0.799+	0.909+	0.064-	0.396+	0.431+
H#	0.000-	0.000-	0.067-	0.963+	0.011-	0.004-	0.021-

Table 7.7: Performance using global processing at the left and right boundaries of each phone when phonemic boundaries are known.

	VOCALIC	CONSON.	HIGH	BACK	LOW	ANTERIOR	CORONAL
IY	0.886+	0.119-	0.740+	0.011-	0.002-	0.057-	0.063-
UW	0.829+	0.340-	0.365+	0.214+	0.000-	0.171-	0.252-
EY	0.927+	0.115-	0.304-	0.017-	0.000-	0.040-	0.052-
OW	0.924+	0.375-	0.004-	0.407+	0.125+	0.218-	0.287-
AA	0.959+	0.123-	0.010-	0.543+	0.359+	0.034-	0.078-
IH	0.871+	0.180-	0.522+	0.020-	0.000-	0.087-	0.113-
UH	0.871+	0.392-	0.271+	0.256+	0.000-	0.155-	0.204-
EH	0.912+	0.211-	0.126-	0.038-	0.065-	0.072-	0.143-
AH	0.903+	0.196-	0.048-	0.460+	0.129-	0.092-	0.161-
AO	0.931+	0.258-	0.018-	0.608+	0.085-	0.107-	0.208-
AE	0.958+	0.113-	0.061-	0.070-	0.163+	0.026-	0.047-
Y	0.614-	0.225-	0.760+	0.004-	0.000-	0.067-	0.169-
W	0.669-	0.396-	0.228+	0.500+	0.007-	0.217-	0.275-
L	0.836+	0.700+	0.060-	0.169-	0.043-	0.504+	0.563+
R	0.911+	0.790+	0.044-	0.063-	0.002-	0.060-	0.714+
N	0.084-	0.955+	0.022-	0.005-	0.000-	0.757+	0.619+
M	0.074-	0.963+	0.004-	0.030-	0.011-	0.830+	0.259-
NG	0.200-	0.848+	0.063+	0.029+	0.025-	0.651-	0.194-
K	0.017-	0.989+	0.485+	0.489+	0.001-	0.320-	0.171-
T	0.013-	0.990+	0.103-	0.019-	0.000-	0.738+	0.545+
P	0.056-	0.974+	0.003-	0.037-	0.005-	0.788+	0.102-
D	0.060-	0.951+	0.039-	0.000-	0.003-	0.627+	0.497+
B	0.165-	0.941+	0.021-	0.024-	0.015-	0.671+	0.106-
G	0.067-	0.972+	0.542+	0.520+	0.011-	0.212-	0.112-
J	0.000-	1.000+	0.475+	0.000-	0.000-	0.415-	0.969+
CH	0.000-	1.000+	0.451+	0.000-	0.000-	0.539-	0.922+
F	0.028-	0.976+	0.023-	0.020-	0.004-	0.808+	0.189-
S	0.003-	0.989+	0.025-	0.000-	0.000-	0.919+	0.885+
SH	0.020-	0.983+	0.334+	0.002-	0.000-	0.481-	0.868+
TH	0.034-	0.980+	0.007-	0.002-	0.005-	0.625+	0.332+
DH	0.115-	0.903+	0.032-	0.025-	0.003-	0.674+	0.469+
Z	0.010-	0.991+	0.026-	0.002-	0.000-	0.940+	0.948+
ZH	0.286-	0.714+	0.286+	0.000-	0.000-	0.000-	0.000+
V	0.333-	0.810+	0.043-	0.025-	0.006-	0.409+	0.241-
H#	0.016-	0.998+	0.017-	0.010-	0.000-	0.105-	0.055-

Table 7.8: Performance using only global processing when phonemic boundaries are unknown.

	ROUND	TENSE	VOICE	CONTIN.	NASAL	STRIDENT	LABIAL
IY	0.009-	0.603+	0.965+	0.972+	0.018-	0.018-	0.002-
UW	0.144+	0.230+	0.973+	0.975+	0.011-	0.005-	0.005-
EY	0.001-	0.350+	0.982+	0.977+	0.018-	0.015-	0.007-
OW	0.205+	0.129+	0.982+	0.972+	0.020-	0.012-	0.007-
AA	0.003-	0.242+	0.981+	0.982+	0.004-	0.005-	0.004-
IH	0.008-	0.143-	0.969+	0.966+	0.017-	0.044-	0.004-
UH	0.142+	0.033-	0.961+	0.965+	0.015-	0.013-	0.002-
EH	0.005-	0.009-	0.977+	0.966+	0.021-	0.017-	0.009-
AH	0.066-	0.045-	0.970+	0.973+	0.021-	0.040-	0.005-
AO	0.316+	0.056-	0.987+	0.991+	0.004-	0.010-	0.004-
AE	0.000-	0.100-	0.978+	0.979+	0.011-	0.008-	0.007-
Y	0.000-	0.270-	0.985+	0.978+	0.026-	0.015-	0.011-
W	0.401+	0.007-	0.985+	0.962+	0.022-	0.001-	0.006-
L	0.091-	0.055-	0.974+	0.946+	0.036-	0.012-	0.006-
R	0.021-	0.012-	0.965+	0.982+	0.006-	0.018-	0.008-
N	0.000-	0.014-	0.937+	0.309-	0.654+	0.029-	0.034-
M	0.003-	0.015-	0.958+	0.290-	0.650+	0.009-	0.347+
NG	0.000-	0.041-	0.981+	0.311-	0.641+	0.000-	0.038-
K	0.006-	0.000-	0.156-	0.132-	0.000-	0.025-	0.025-
T	0.000-	0.001-	0.178-	0.170-	0.002-	0.142-	0.024-
P	0.012-	0.009-	0.310-	0.192-	0.000-	0.047-	0.521+
D	0.000-	0.010-	0.650+	0.282-	0.021-	0.127-	0.067-
B	0.012-	0.032-	0.791+	0.376-	0.009-	0.000-	0.409+
G	0.000-	0.006-	0.592+	0.140-	0.011-	0.028-	0.034-
J	0.000-	0.000-	0.542+	0.206-	0.000-	0.602+	0.000-
CH	0.000-	0.000-	0.092-	0.208-	0.000-	0.683+	0.000-
F	0.007-	0.003-	0.146-	0.929+	0.003-	0.667+	0.464+
S	0.000-	0.001-	0.318-	0.951+	0.000-	0.900+	0.022-
SH	0.000-	0.001-	0.155-	0.886+	0.000-	0.846+	0.055-
TH	0.000-	0.002-	0.120-	0.818+	0.005-	0.268-	0.141-
DH	0.018-	0.010-	0.624+	0.756+	0.085-	0.097-	0.019-
Z	0.000-	0.002-	0.837+	0.965+	0.004-	0.922+	0.000-
ZH	0.000-	0.000-	0.286+	1.000+	0.000-	0.000+	0.000-
V	0.006-	0.039-	0.729+	0.921+	0.060-	0.195+	0.138+
H#	0.001-	0.003-	0.102-	0.924+	0.021-	0.039-	0.028-

Table 7.9: Performance using only global processing when phonemic boundaries are unknown.

7.3 Individual results – Combined Local and Global Processing

Shown in the tables of this section are results of linguistic feature estimation for individual phonemes, combining the global processing of Chapter 4 with the local processing of Chapter 3 by averaging the probability estimates derived from each method.

Tables 7.10 and 7.11 reflect performance in the case of known TIMIT boundaries, while tables 7.12 and 7.13 indicate the performance in the case of unknown TIMIT markings.

	VOCALIC	CONSON.	HIGH	BACK	LOW	ANTERIOR	CORONAL
IY	0.962+	0.047-	0.955+	0.000-	0.000-	0.015-	0.002-
UW	0.995+	0.288-	0.534+	0.421+	0.000-	0.151-	0.200-
EY	0.998+	0.003-	0.362-	0.000-	0.000-	0.002-	0.002-
OW	0.998+	0.046-	0.017-	0.727+	0.568+	0.113-	0.129-
AA	0.999+	0.012-	0.001-	0.865+	0.813+	0.001-	0.005-
IH	0.984+	0.029-	0.713+	0.013-	0.000-	0.003-	0.016-
UH	0.993+	0.199-	0.269+	0.256+	0.000-	0.101-	0.144-
EH	0.997+	0.057-	0.084-	0.057-	0.154-	0.003-	0.021-
AH	0.996+	0.075-	0.022-	0.557+	0.169-	0.008-	0.040-
AO	0.982+	0.057-	0.007-	0.939+	0.207-	0.001-	0.015-
AE	0.998+	0.017-	0.038-	0.021-	0.713+	0.001-	0.001-
Y	0.588-	0.139-	0.828+	0.000-	0.000-	0.004-	0.120-
W	0.317-	0.211-	0.645+	0.776+	0.001-	0.068-	0.044-
L	0.862+	0.788+	0.041-	0.160-	0.024-	0.712+	0.658+
R	0.975+	0.841+	0.066-	0.021-	0.002-	0.005-	0.787+
N	0.046-	0.977+	0.008-	0.003-	0.004-	0.832+	0.621+
M	0.028-	0.994+	0.000-	0.003-	0.004-	0.904+	0.064-
NG	0.130-	0.870+	0.213+	0.003+	0.006-	0.660-	0.308-
K	0.000-	1.000+	0.742+	0.592+	0.000-	0.135-	0.088-
T	0.000-	1.000+	0.168-	0.000-	0.000-	0.756+	0.679+
P	0.000-	1.000+	0.009-	0.000-	0.000-	0.828+	0.014-
D	0.000-	0.992+	0.194-	0.013-	0.000-	0.671+	0.642+
B	0.071-	0.994+	0.021-	0.018-	0.000-	0.621+	0.044-
G	0.000-	1.000+	0.659+	0.564+	0.000-	0.201-	0.011-
J	0.000-	1.000+	0.946+	0.000-	0.000-	0.025-	0.969+
CH	0.000-	1.000+	0.908+	0.000-	0.000-	0.085-	0.993+
F	0.000-	1.000+	0.029-	0.015-	0.000-	0.842+	0.082-
S	0.001-	0.999+	0.031-	0.000-	0.000-	0.981+	0.988+
SH	0.002-	0.998+	0.819+	0.000-	0.000-	0.178-	0.988+
TH	0.002-	0.998+	0.048-	0.002-	0.000-	0.559+	0.359+
DH	0.032-	1.000+	0.026-	0.007-	0.000-	0.784+	0.457+
Z	0.002-	0.999+	0.056-	0.000-	0.000-	0.940+	0.997+
ZH	0.000-	1.000+	0.000+	0.000-	0.000-	1.000-	1.000+
V	0.109-	0.941+	0.052-	0.003-	0.001-	0.616+	0.109-
H#	0.003-	0.998+	0.001-	0.000-	0.000-	0.013-	0.004-

Table 7.10: Performance combining local processing with global processing when phonemic boundaries are known. The local estimate derives from dependent modeling of anterior-coronal and high-back-low.

	ROUND	TENSE	VOICE	CONTIN.	NASAL	STRIDENT	LABIAL
IY	0.000-	0.954+	0.998+	0.980+	0.018-	0.001-	0.001-
UW	0.649+	0.529+	0.998+	0.995+	0.002-	0.000-	0.005-
EY	0.000-	0.778+	0.999+	0.999+	0.001-	0.001-	0.001-
OW	0.817+	0.426+	1.000+	0.998+	0.001-	0.000-	0.001-
AA	0.330-	0.499+	0.999+	0.999+	0.000-	0.001-	0.000-
IH	0.045-	0.347-	0.999+	0.998+	0.002-	0.001-	0.001-
UH	0.396+	0.171-	1.000+	0.998+	0.000-	0.000-	0.004-
EH	0.078-	0.062-	0.999+	0.998+	0.001-	0.000-	0.001-
AH	0.209-	0.105-	0.999+	0.997+	0.002-	0.000-	0.003-
AO	0.887+	0.173-	0.998+	0.999+	0.000-	0.001-	0.001-
AE	0.000-	0.063-	0.999+	0.999+	0.000-	0.000-	0.001-
Y	0.000-	0.554-	0.985+	0.996+	0.004-	0.000-	0.004-
W	0.906+	0.007-	0.990+	0.995+	0.000-	0.000-	0.004-
L	0.461-	0.142-	0.998+	0.968+	0.020-	0.000-	0.014-
R	0.032-	0.039-	0.991+	0.997+	0.000-	0.000-	0.013-
N	0.001-	0.019-	0.996+	0.171-	0.795+	0.004-	0.116-
M	0.011-	0.007-	0.976+	0.136-	0.866+	0.002-	0.818+
NG	0.000-	0.159-	1.000+	0.146-	0.819+	0.000-	0.086-
K	0.000-	0.000-	0.068-	0.050-	0.000-	0.017-	0.064-
T	0.000-	0.000-	0.095-	0.106-	0.000-	0.115-	0.016-
P	0.000-	0.000-	0.161-	0.099-	0.000-	0.000-	0.905+
D	0.000-	0.010-	0.663+	0.127-	0.000-	0.104-	0.054-
B	0.068-	0.015-	0.882+	0.368-	0.003-	0.000-	0.706+
G	0.000-	0.000-	0.564+	0.039-	0.000-	0.011-	0.045-
J	0.000-	0.000-	0.460+	0.028-	0.000-	0.907+	0.000-
CH	0.000-	0.000-	0.000-	0.195-	0.000-	0.993+	0.000-
F	0.000-	0.000-	0.006-	0.966+	0.000-	0.727+	0.898+
S	0.000-	0.000-	0.202-	0.992+	0.000-	0.997+	0.031-
SH	0.000-	0.001-	0.012-	0.957+	0.000-	0.988+	0.117-
TH	0.002-	0.000-	0.025-	0.900+	0.000-	0.323-	0.345-
DH	0.000-	0.009-	0.712+	0.657+	0.141-	0.018-	0.115-
Z	0.000-	0.000-	0.738+	0.981+	0.001-	0.995+	0.012-
ZH	0.000-	0.000-	1.000+	1.000+	0.000-	1.000+	0.000-
V	0.020-	0.017-	0.838+	0.914+	0.038-	0.346+	0.752+
H#	0.001-	0.001-	0.083-	0.990+	0.004-	0.003-	0.020-

Table 7.11: Performance combining local processing with global processing when phonemic boundaries are known. The local estimate derives from dependent modeling of anterior-coronal and high-back-low.

	VOCALIC	CONSON.	HIGH	BACK	LOW	ANTERIOR	CORONAL
IY	0.898+	0.127-	0.878+	0.014-	0.003-	0.050-	0.045-
UW	0.858+	0.336-	0.579+	0.457+	0.002-	0.185-	0.194-
EY	0.954+	0.068-	0.370-	0.004-	0.008-	0.033-	0.036-
OW	0.940+	0.267-	0.028-	0.611+	0.406+	0.235-	0.244-
AA	0.964+	0.080-	0.011-	0.771+	0.621+	0.022-	0.056-
IH	0.884+	0.132-	0.687+	0.024-	0.026-	0.070-	0.093-
UH	0.862+	0.383-	0.319+	0.254+	0.007-	0.118-	0.193-
EH	0.921+	0.128-	0.119-	0.058-	0.168-	0.062-	0.092-
AH	0.914+	0.176-	0.057-	0.542+	0.189-	0.079-	0.117-
AO	0.952+	0.143-	0.045-	0.835+	0.192-	0.071-	0.122-
AE	0.960+	0.044-	0.041-	0.019-	0.676+	0.020-	0.016-
Y	0.584-	0.161-	0.876+	0.000-	0.000-	0.019-	0.124-
W	0.412-	0.212-	0.617+	0.762+	0.011-	0.093-	0.096-
L	0.846+	0.656+	0.084-	0.250-	0.059-	0.542+	0.561+
R	0.921+	0.778+	0.089-	0.057-	0.019-	0.036-	0.713+
N	0.078-	0.944+	0.037-	0.004-	0.004-	0.820+	0.650+
M	0.067-	0.946+	0.015-	0.021-	0.016-	0.861+	0.116-
NG	0.200-	0.790+	0.235+	0.010+	0.022-	0.641-	0.333-
K	0.013-	0.988+	0.702+	0.541+	0.002-	0.152-	0.108-
T	0.013-	0.989+	0.209-	0.005-	0.001-	0.701+	0.725+
P	0.050-	0.971+	0.045-	0.017-	0.005-	0.744+	0.060-
D	0.052-	0.961+	0.166-	0.026-	0.000-	0.585+	0.557+
B	0.141-	0.929+	0.059-	0.038-	0.009-	0.615+	0.094-
G	0.067-	0.972+	0.693+	0.564+	0.006-	0.123-	0.089-
J	0.000-	1.000+	0.944+	0.000-	0.000-	0.056-	0.975+
CH	0.000-	1.000+	0.802+	0.000-	0.000-	0.242-	0.983+
F	0.032-	0.974+	0.042-	0.031-	0.005-	0.809+	0.094-
S	0.003-	0.998+	0.050-	0.000-	0.000-	0.946+	0.969+
SH	0.020-	0.983+	0.788+	0.001-	0.000-	0.269-	0.930+
TH	0.034-	0.970+	0.027-	0.007-	0.007-	0.618+	0.316+
DH	0.116-	0.906+	0.062-	0.012-	0.003-	0.707+	0.488+
Z	0.011-	0.991+	0.068-	0.001-	0.000-	0.912+	0.965+
ZH	0.286-	0.714+	0.000+	0.000-	0.000-	0.286-	0.286+
V	0.342-	0.778+	0.113-	0.078-	0.013-	0.477+	0.187-
H#	0.016-	0.993+	0.021-	0.008-	0.001-	0.069-	0.045-

Table 7.12: Performance combining local processing with global processing when phonemic boundaries are unknown. The local estimate derives from dependent modeling of high-back-low and anterior-coronal.

	ROUND	TENSE	VOICE	CONTIN.	NASAL	STRIDENT	LABIAL
IY	0.021-	0.876+	0.966+	0.967+	0.025-	0.017-	0.005-
UW	0.610+	0.505+	0.975+	0.964+	0.020-	0.005-	0.061-
EY	0.010-	0.682+	0.983+	0.979+	0.018-	0.014-	0.014-
OW	0.809+	0.444+	0.986+	0.969+	0.025-	0.015-	0.026-
AA	0.335-	0.403+	0.980+	0.982+	0.004-	0.005-	0.027-
IH	0.051-	0.298-	0.984+	0.962+	0.017-	0.047-	0.019-
UH	0.431+	0.147-	0.974+	0.963+	0.013-	0.018-	0.022-
EH	0.071-	0.041-	0.979+	0.960+	0.029-	0.017-	0.025-
AH	0.332-	0.131-	0.980+	0.971+	0.024-	0.040-	0.030-
AO	0.793+	0.211-	0.986+	0.988+	0.006-	0.010-	0.014-
AE	0.013-	0.104-	0.981+	0.982+	0.011-	0.007-	0.011-
Y	0.000-	0.573-	0.996+	0.955+	0.022-	0.019-	0.019-
W	0.848+	0.054-	0.989+	0.970+	0.023-	0.005-	0.053-
L	0.488-	0.209-	0.978+	0.945+	0.043-	0.008-	0.026-
R	0.067-	0.051-	0.970+	0.978+	0.009-	0.018-	0.031-
N	0.006-	0.025-	0.970+	0.198-	0.779+	0.028-	0.129-
M	0.044-	0.014-	0.961+	0.162-	0.833+	0.009-	0.759+
NG	0.035-	0.127-	0.981+	0.254-	0.740+	0.000-	0.067-
K	0.009-	0.005-	0.099-	0.117-	0.000-	0.027-	0.086-
T	0.001-	0.003-	0.144-	0.164-	0.002-	0.166-	0.045-
P	0.025-	0.011-	0.209-	0.174-	0.000-	0.045-	0.736+
D	0.000-	0.021-	0.642+	0.246-	0.016-	0.135-	0.101-
B	0.091-	0.068-	0.868+	0.359-	0.026-	0.009-	0.703+
G	0.006-	0.011-	0.564+	0.117-	0.011-	0.022-	0.022-
J	0.000-	0.000-	0.540+	0.040-	0.000-	0.895+	0.000-
CH	0.000-	0.000-	0.038-	0.188-	0.000-	0.860+	0.034-
F	0.016-	0.011-	0.070-	0.923+	0.002-	0.681+	0.816+
S	0.001-	0.001-	0.239-	0.979+	0.000-	0.969+	0.028-
SH	0.001-	0.004-	0.061-	0.918+	0.000-	0.972+	0.094-
TH	0.007-	0.007-	0.193-	0.834+	0.005-	0.216-	0.293-
DH	0.003-	0.032-	0.699+	0.719+	0.118-	0.084-	0.122-
Z	0.000-	0.009-	0.810+	0.962+	0.005-	0.967+	0.012-
ZH	0.000-	0.000-	1.000+	1.000+	0.000-	0.286+	0.000-
V	0.096-	0.060-	0.762+	0.908+	0.074-	0.215+	0.510+
H#	0.004-	0.006-	0.115-	0.947+	0.019-	0.031-	0.048-

Table 7.13: Performance combining local processing with global processing when phonemic boundaries are unknown. The local estimate derives from dependent modeling of high-back-low and anterior-coronal.

Bibliography

- [1] Agard, F. and R. Di Pietro. *The Sounds of English and Italian*. Chicago: The University of Chicago Press. 1965.
- [2] Anderson, T.W. and H. Rubin. "Statistical Inference in Factor Analysis." *Proc. of the Third Berkeley Symposium on Mathematical Statistics and Probability*. 1956. Vol. 5. pp. 111-150.
- [3] Chernoff, H. Unpublished note. 1977.
- [4] Chomsky, N. and M. Halle. *Sound Pattern of English*. New York: Harper & Row. 1968.
- [5] Clements, G. "The Geometry of Phonological Features." in *Phonology Yearbook 2*. 1985. pp 225-252.
- [6] Croxton, F. *Elementary Statistics with Applications in Medicine and the Biological Sciences*. New York: Dover Publications, Inc. 1953.
- [7] Deng, L. and K. Erler. "Structural Design of Hidden Markov Model Speech Recognizer Using Multivalued Phonetic Features: Comparison with Segmental Speech Units." *Journal of the Acoustical Society of America*. Vol. 92 No. 6. Dec. 1992. pp. 3058-3067.
- [8] Digilakis, V., "Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition." PhD Thesis, Boston University. 1992.
- [9] Fant, G. *Speech Sounds and Features*. Cambridge, MA: MIT Press. 1973.

- [10] Gish, H. "Identification of Speakers Engaged in Dialog." in *Proc. of IEEE ICASSP*. 1993. p. II-383.
- [11] Gish, H. Personal Communication. 1993.
- [12] Lee, K. and H. Hon. "Speaker Independent Phoneme Recognition Using Hidden Markov Models." *IEEE Trans. ASSP* Vol. 37, No. 11. Nov. 1989. pp 1641-1648.
- [13] Lee, K. *Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. PhD Thesis, Computer Science Department, Carnegie-Mellon University. 1988.
- [14] Leung, H. and V. Zue. "Phonetic Classification Using Multi-Layer Perceptrons." *Proc. of IEEE ICASSP*. 1990. pp. 525-528.
- [15] McCarthy, J. "Feature Geometry and Dependency: A Review." in *Phonetica* 43. 1988; 45:84-108.
- [16] Meng, H. *The Use of Distinctive Features For Automatic Speech Recognition*. S.M. Thesis, Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology. 1991.
- [17] Mumford, D. *Pattern Theory: A Unifying Perspective*. CICS Report.
- [18] Papoulis, A. *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill. 1984.
- [19] Pickles, J. *An Introduction to the Physiology of Hearing*. London: Academic Press. 1982.
- [20] Potter, R., G. Kopp, and H. Kopp. *Visible Speech*. New York: Dover Publications. 1966.
- [21] Rabiner, L. and R. Schafer. *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall. 1987.

- [22] Sagey, E. *The Representation of Features and Relations in Non-linear Phonology*. PhD Dissertation. Massachusetts Institute of Technology. 1986.
- [23] Schwartz, R. and Y. Chow. "The N-Best Algorithm: An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses." *Proc. of IEEE ICASSP*. 1990. pp. 81-84.
- [24] Stevens, K. Personal communication. 1993.
- [25] Stevens, K. "Evidence for the Role of Acoustic Boundaries in the Perception of Speech Sounds" in *Phonetic Linguistics. Essays in Honor of Peter Ladefoged*. V. Fromkin, ed. Orlando, FL: Academic Press, Inc. 1985.
- [26] Stevens, K. "Phonetic Features and Lexical Access." Presented at em Symposium on Advanced Man-Machine Interface Through Spoken Language. November 19-22, 1988. Hawaii.
- [27] Stevens, K. "On The Quantal Nature of Speech." in *Jornal of Phonetics*. Vol. 17, pp 3-45. 1989.
- [28] Stevens, K. and M. Halle. "Remarks On Analysis By Synthesis and Distinctive Features." in *Models for the Perception of Speech and Visual Form*. Cambridge, MA: The MIT Press. 1964.
- [29] Waibel, A., et. al. "Phoneme Recognition Using Time-Delay Neural Networks." *IEEE Trans. ASSP*. Vol. 37, No. 3. March, 1989.
- [30] Zue, V., S. Seneff, and J. Glass. "Speech Database Development: TIMIT and Beyond," Presented at Workshop on Speech Input/Output Assessment and Speech Databases, Amsterdam, the Netherlands. September, 1989.